Identification, estimation and control of gene expression and metabolic network dynamics

Eugenio Cinquemani

Université Grenoble-Alpes
École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique

Habilitation à Diriger les Recherches
Specialité: Mathématiques appliquées et informatique

Mémoire présenté par
Eugenio CINQUEMANI
Inria Grenoble – Rhône-Alpes

Identification, estimation and control of gene expression and metabolic network dynamics

---

Soutenue le 20 Novembre 2019 devant un jury composé de

Adeline Leclercq Samson	Président	Université Grenoble-Alpes
Béatrice Laroche	Rapporteur	INRA Jouy-en-Josas
Julio Banga	Rapporteur	IIM-CSIC, Vigo (Spain)
Madalena Chaves	Rapporteur	Inria Sophia Antipolis – Mediterranée
Steffen Waldherr	Examinateur	KU Leuven (Belgium)
Hidde de Jong	Examinateur	Inria Grenoble – Rhône-Alpes

Abstract

The great technological advances of experimental biology have enabled the study of microbial life at an unprecedented level of detail, opening up new opportunities but also new challenges concerning the exploitation of the data and the mathematical modelling of biological systems. Systems biology has been intensely developed trying to profit at most from these new opportunities and to provide answers to the new questions. Advances in experimental technology and synthetic biology have also started to turn biology into an engineering discipline in addition to its traditional status of explanatory science. A noteworthy frontier is the automatic control of synthetically engineered cellular systems. The interdisciplinary nature of these activites has attracted the attention of many researchers of diverse background, in particular, from the automatic control community.

In this manuscript, I present an overview of my research at the intersection of systems biology, statistical estimation and control theory. The research presented has been developed for the largest part at IBIS, the Inria project-team that I am member of, also thanks to supervision of Ph.D students and other young researchers. With reference to selected papers among my publications, I discuss several challenges in identification, analysis, estimation and control of microbial systems. Developed in connection with real systems and data, these contributions also witness the potential of this application to stimulate methodological research of broader relevance.

In Chapter 1 we consider reconstruction of two types of deterministic cellular regulation models from population-mean data. In the context of gene regulatory networks, based on the so-called sign pattern approach to network inference, we focus on an application to time-course *E.coli* gene expression data. With reference to [107], in addition to the effectiveness of the reconstruction method, we discuss the importance of accounting for often neglected details in the treatment of time-course data. Next, in the context of metabolism regulatory network identification, we concentrate on the fundamental problem of identifiability. In the special case of pseudo-linear regulatory models, with reference to [10], we discuss structural and practical identifiability and related model reduction methods, illustrating the relevance of these problems on a real dataset from the *E.coli* central carbon metabolism.

In Chapter 2 we focus on the modelling of cell-to-cell variability in isogenic populations. With reference to [67], we first discuss the Mixed-Effects (ME) approach to the modelling of gene expression response in isogenic cell populations. In this approach, an identical dynamical model for all cells takes statistically independent parameter values across individuals, following the laws of a common population distribution. With reference to [72], we describe an extension of the ME approach called Auto-Regressive ME (ARME). This approach accounts for parental relationships among individuals and allows one to exploit lineage information to quantify single-cell parameter inheritance and variability at cell division. Results on a common yeast osmotic shock gene expression response dataset demonstrate the ability of ME to capture the variability of gene expression dynamics across cells, and that of ARME to additionally quantify how this

variability is built up along cellular generations.

In Chapter 3 we consider two different dynamical process estimation problems. The first concerns the reconstruction of microbial growth and substrate exchange rates from metabolomics time-course data. Here, with reference to [24], we discuss a Gaussian process formulation of regularized estimation that allows for the accurate reconstruction of varying metabolic regimes and their transitions in changing environments, and its demonstration on data from several bacterial species grown in batch or fed-batch. The second addresses inference of promoter activation statistics from population-snapshot gene expression data. With reference to [22], we discuss a generalization of the so-called moment equations for a class of reaction networks with stochastic rate fluctuations, and their application to the processing of fluorescent reporter protein measurements in single-cells, showing the possibility to reconstruct statistics of gene activation (notably the autocovariance function) from fluorescence mean and variance profiles.

In Chapter 4 we discuss microbial control in two different respects. The first is the investigation of the policies that microorganisms naturally put in place to optimize growth, and how these can be artificially changed to optimize other objectives, such as production of molecules of interest. With reference to [25], we discuss in particular the effect of accounting for costs involved in rapidly changing control actions, and of the time horizon over which the problem is considered. The second is the feedback control of microbial communities, with focus on synthetically engineered communities. As a preview of a publication in preparation [73], we discuss the interest of the problem in terms of coexistence and interaction dynamics of different species as well as productivity enhancement of target molecules. We conclude the chapter with a detour into control of stochastic dynamical systems, a subject of potential interest in biology for single-cell, stochastic model-based control. With reference to [23], we discuss problems with probabilistic constraints, and means to guarantee convexity, whence practicality of the resulting optimization problems.

Finally, in Chapter 5, we draw conclusions about the research presented. We discuss ongoing activities in the context of current projects and possible perspectives for further developments. More space is dedicated here to the research project that I coordinate toward automated control of synthetic microbial communities, whose scientific and technogical developments are expected to shape my own research activity and that of IBIS in the upcoming years.

Résumé en Français

Les grandes avancées technologiques en biologie expérimentale permettent d'étudier la vie microbienne à un niveau de détail sans précédent, ouvrant la voie à de nouvelles possibilités mais également de nouveaux défis concernant l'exploitation des données et la modélisation mathématique des systèmes biologiques. La biologie des systèmes a été intensément développée pour exploiter au mieux ces nouvelles opportunités et pour fournir des réponses à des questions nouvelles. Les progrès de la technologie expérimentale et de la biologie synthétique ont également commencé à transformer la biologie en une discipline d'ingénierie, en plus de son statut traditionnel de science explicative. À la frontière de ces développements, on trouve notamment le contrôle automatique des systèmes cellulaires synthétiques. La nature interdisciplinaire de ces activités a attiré l'attention de nombreux chercheurs de différents domaines, en particulier de la communauté de contrôle automatique.

Dans ce manuscrit, je présente un aperçu de mes recherches à l'intersection de la biologie des systèmes, de l'estimation statistique et de la théorie du contrôle. Ces travaux ont été développés pour la plupart au sein de l'équipe-projet Inria IBIS dont je suis membre, grâce en partie à l'encadrement de doctorants et d'autres jeunes chercheurs. À partir d'une sélection d'articles parmi mes publications, je traite de problèmes d'identification, d'analyse, d'estimation et de contrôle des systèmes microbiens. Développés en relation avec des systèmes et des données réels, ces contributions témoignent également du potentiel de ces applications pour stimuler le développement d'outils mathématiques d'intérêt plus général.

Dans le Chapitre 1, nous considèrerons la reconstruction de deux types de modèles déterministes de régulation cellulaire à partir de données moyennes de population. Dans le contexte des réseaux de régulation des gènes, à partir d'une approche dite des "sign patterns" pour l'inférence de réseau, nous nous concentrerons sur une application à des séries temporelles d'expression génique pour la bactérie *E.coli*. En se basant sur [107], outre l'efficacité de la méthode de reconstruction, nous discuterons de l'importance de prendre en compte des détails souvent négligés dans le traitement des données temporelles. Ensuite, dans le contexte de l'identification des réseaux de régulation métabolique, nous nous focaliserons sur le problème fondamental de l'identifiabilité. Dans le cas particulier des modèles de régulation pseudo-linéaires, en référence à [10], nous discuterons de l'identifiabilité structurelle et pratique, des méthodes de réduction de modèle associées, et nous illustrerons la pertinence de ces problèmes sur un jeu de données réelles sur le métabolisme du carbone chez *E.coli*.

Dans le Chapitre 2, nous nous intéresserons à la modélisation de la variabilité intercellulaire dans des populations isogéniques. Sur la base de [67], nous présenterons d'abord l'approche dite "Mixed Effects" (ME) pour la modélisation de l'expression génique. Dans cette approche, un modèle dynamique identique pour toutes les cellules prend des valeurs de paramètres statistiquement indépendantes entre les individus, conformément à une distribution de probabilité commune à la population. En référence à [72], nous présenterons une extension de l'approche ME nommée "Auto-Regressive" ME (ARME).

Cette approche prend en compte et exploite les relations parentales entre les individus pour quantifier l'héritabilité et la variabilité des paramètres d'expression génique lors de la division cellulaire. Les résultats sur des données de réponse génique de la levure à des chocs osmotiques démontrent la capacité de la méthode ME à capturer la variabilité intercellulaire des dynamiques d'expression génique, et celle de la méthode ARME à quantifier aussi la manière dont cette variabilité se construit le long des générations cellulaires.

Au Chapitre 3, nous examinerons deux problèmes différents d'estimation de processus dynamiques microbiens. Le premier concerne la reconstruction des taux de croissance et d'échange des substrats avec l'environnement à partir de données métabolomiques temporelles. En référence à [24], nous présenterons une formulation régularisée du problème d'estimation basée sur les techniques de processus Gaussiens. Appliquée à des jeux de données de plusieurs espèces bactériennes cultivées en batch ou fed-batch, cette méthode permet la reconstruction précise de divers régimes métaboliques et de leurs transitions dans des environnements dynamiques. Le deuxième problème porte sur l'inférence des statistiques d'activation d'un promoteur à partir des instantanées du niveau d'expression génique dans une population cellulaire. En se basant sur [22], nous discuterons d'une généralisation des équations de moments pour une classe de réseaux de réactions avec des taux stochastiques, et de leur application au traitement de données de rapporteur fluorescent d'expression génique dans des cellules individuelles. Nous montrerons la possibilité de reconstruire les statistiques d'activation des gènes (notamment la fonction d'auto-covariance) à partir des séries temporelles de moyenne et variance de la fluorescence.

Au Chapitre 4, nous aborderons le contrôle microbien à deux niveaux différents. Le premier concerne l'étude des stratégies naturelles de contrôle mises en place par les micro-organismes afin d'optimiser la croissance, et leur modification artificielle pour optimiser des objectifs différents, tels que la production de bio-molécules d'intérêt. En référence à [25], nous discuterons notamment des coûts liés aux variations rapides de l'action de contrôle et des effets de l'horizon temporel sur lequel le problème est considéré. Le deuxième niveau de contrôle est le contrôle en rétroaction des communautés microbiennes synthétiques. Anticipant une publication en préparation [73], nous discuterons de l'intérêt du problème en termes de coexistence et de dynamique d'interaction de différentes espèces ainsi que d'amélioration de la productivité des molécules cibles. Nous terminerons le chapitre par un détour par le contrôle des systèmes dynamiques stochastiques, sujet d'intérêt en biologie pour le contrôle des cellules individuelles à partir de modèles stochastiques. En référence à [23], nous discuterons des problèmes à contraintes probabilistes, de leur convexité et des méthodes d'approximation convexe, un prérequis important pour leur analyse et résolution numérique.

Le Chapitre 5 est dédié à une discussion finale sur les recherches présentées, les activités de recherche dans des projets en cours et les perspectives de développement ultérieur. Une part importante est notamment consacrée au projet de recherche que je coordonne sur le contrôle automatisé de communautés microbiennes synthétiques, dont les développements scientifiques et techniques vont façonner ma propre activité de recherche et celle d'IBIS dans les années à venir.

Contents

Introduction

The great technological advances that experimental biology has witnessed in the last decades have enabled the study of microbial life at an unprecedented detail level. Combined with the use of microfluidic devices, techniques such as fluorescence videomicroscopy allow one to monitor gene expression levels quantitatively over time even at single-cell resolution [110]. Advanced metabolomics techniques have made it possible to precisely monitor the concentration of extracellular metabolites over time-course experiments [86]. These and other developments have opened up new opportunities but also new challenges concerning the exploitation of the data and the mathematical modelling of biological systems.

Systems biology, which can be pictured as the combined experimental, mathematical and computational study of biology from the perspective of system theory, has been intensely developed in parallel, trying to profit at most from these new opportunities and to provide answers to the new questions. Quite naturally, this interdisciplinary activity has stimulated biologists to open up to novel methodologies and, conversely, has attracted the attention of applied mathematics and research from several mathematics-related disciplines in search of original applications. In particular, it has attracted many researchers from the automatic control community, taking the development of dynamical systems & control theory toward new frontiers [15, 63, 62, 85, 1, 3, 126].

The advances in experimental technology and synthetic biology have also started to turn biology into an engineering discipline in addition to its traditional status of explanatory science. A noteworthy frontier of laboratory research is the computer-driven feedback control of synthetically engineered cellular systems. Today, as demonstrated by several recent contributions, real-time computer-driven control of suitably designed microbial populations allows one to steer cellular population toward predefined objectives even at the level of single-cells [102, 76, 114, 77, 91]. Besides the interest as a means of exploration of cellular dynamics, the ability to real-time control cells opens up intriguing research avenues such as, for instance, the prototyping of synthetic circuits via computer simulations connected to real biological systems [18].

Provided nontrivial scaling-up to real-world scenarios, laboratory exploration of real-time control has potential applications to challenges of great societal interest, such as, in particular, biotechnological production of valuable molecules. Nowadays, reengineering of microbial species to the purpose of building artificial cell factories is being addressed also from the perspective of optimal control [44, 125]. A booming research direction in

this context is the synthesis of artificial microbial communities [7, 104]. Leveraging the concept that microbial consortia can outperform single species in the accomplishment of profitable tasks, detailed investigation of microbial community interaction dynamics is being pursued at a theoretical as well as at an experimental level [122, 46, 108].

Addressing the challenges coming from this new twist of biology largely profits from specific competences at the intersection between systems biology, dynamical systems and control theory [116]. Processing new forms of data from the ever-developing experimental techniques necessitates new modelling and inference methods to precisely capture the biologically relevant information carried by the data. A chief example of the importance of these tools in today's biology is the Kalman filter [56], a control-theoretic tool that is imposing itself on biological data processing [64, 24, 116]. Real-time control applications obviously require control-theoretic competences for the suitable modelling of the biological systems and the design of model-based (optimal) control strategies, *e.g.* Model-Predictive Control (MPC) [114]. The problems addressed by these tools are not only of interest to biology and its technological applications. They also lead to the development of theoretical and computational tools that can find applications in other scientific domains, and stimulate new research avenues in applied mathematics, control theory, and all the contributing fields [48, 68].

Since the early stages of my research career, based on a control-theoretic and statistics background, I have been investigating stochastic estimation and control problems. In my Master thesis, I addressed Bayesian deconvolution problems based on multiresolution analysis techniques. Still at the University of Padova, during my Ph.D. in the Automatic Control and System Identification group I kept working on inference problems, focusing on theoretical dynamical estimation problems for switching linear Markovian processes. Inspiring research being developed around that time, where switching dynamical systems were proposed as a description of genetic regulatory systems, captured my interest toward systems biology. In my PostDoc at the Automatic Control Lab of ETH Zurich, I focused on systems biology research topics, in particular, gene regulatory network inference from time-course gene expression data. In parallel, I developed a second research line on control of stochastic dynamical systems via a Model Predictive Control (MPC) approach. The increasing interest and opportunities in systems biology eventually brought me to join IBIS, an Inria systems biology project-team [51], as a permanent researcher in 2009.

In this manuscript I present an overview of my recent and current research, which stands at the intersection of systems biology, statistical estimation and control theory. In the selected papers discussed in the manuscript, I investigate several of the methodological challenges that have been outlined above in connection with real microbial systems and data. In line with previous reflections, the contributions provided in these papers are of interest not only to systems biology, but they also address original problems in the fields of estimation and dynamical system analysis and control. An example of a control-theoretic contribution that may well apply to problems of specific relevance for systems biology is also reported. The research described in this work also provides the starting point of my future research directions, both for systems biology and for possible

relevant developments in related fields.

The manuscript is organized as follows. In Chapter 1, I discuss reconstruction of genetic and metabolic networks, respectively, from quantitative gene expression dynamics and metabolic flux data. Special attention is devoted to identifiability analysis, a subject of pressing importance in the quantitative modelling of cellular regulatory dynamics from experimetal data. The two papers reported are the conclusive part of methodological developments initiated in previous work and concisely reviewed. In Chapter 2, questions around the modelling of individual-cell gene expression variability are addressed with reference to real gene expression osmotic shock response data in yeast. The two papers represent one the evolution of the other, taking the initially developed models of cell response variability into the context of modelling and experimental observation of gene expression along lineage trees. In Chapter 3, focus is shifted to estimation of unknown time-dependent quantities from time-course data. In a first work, focus is on Kalman smoothing techniques and their adaptation to dynamical models of microbial growth, for the precise estimation of growth and substrate exchange rates from time-course (population-average) metabolomics data. In a second work, the problem of resconstructing properties of the unknown stochastic dynamics of gene activation from fluorescent reporter population-snapshot data is considered instead. In Chapter 4, attention goes to control of microbial systems. Control is investigated at two different levels, on the one hand, natural and artificial optimal resource allocation problems at an intracellular level, and on the other hand, modelling and analysis of synthetic microbial communities toward control for optimal product biosynthesis. Due to the novelty of this research line, a preliminary conference paper and a publication in preparation are discussed for this part. The chapter is concluded with the discussion of a theoretic contribution on stochastic optimal control of dynamical systems with constraints. In all of these chapters, bibliographic reference and abstract of the selected papers are reported. Finally, Chapter 5 draws conclusions about the reported research, together with additional discussion of ongoing projects that I coordinate or participate in, and an outlook of my upcoming research activities.

Chapter 1

Identification of deterministic metabolic and gene network models

1.1 Identification of quantitative network models in biology

Identification of regulatory interactions in cells is a crucial challenge since before the advent of systems biology. The discovery of regulatory mechanisms is key not only to the understanding of how cells and more complex organisms sustain life in response to changing environments, but it has become a central challenge for several applications. Chiefly among these is new generation medicine, where therapies based on genome editing for enforcing suitable regulation are sought [6, 69].

Great advances in modern experimental techniques, along with a novel viewpoint that emerged in response to these advances, has radically modified the approach to regulatory network discovery. Identification of regulatory interactions has moved in a range of a few years from standard statistical approaches such as correlation analysis of gene expression on-off data, to challenges such as reconstruction of quantitative, dynamical gene expression regulatory interaction models from time-course gene expression profiles [20, 117].

The modern network reconstruction endeavor has stimulated the development of a variety of modelling frameworks and network reconstruction techniques. Complicated by the ambition to establish approaches of some general applicability, the importance of the quest for appropriate methodologies is witnessed by a yearly international competition dedicated to it [36, 71]. However, the development of network estimation algorithms turns out to be only a small part of the problem. Hidden behind the problem is a number of conceptual and practical issues, in particular, the well-posedness of the estimation problems and the informativity and nature of the data.

The appropriateness of the reconstruction problem statement revolves around the existence and uniqueness of the solution, that is, a problem of identifiability [4, 19, 99].

Essentially a structural property of the model class in the system identification theory, the problem becomes more involved, and sometimes vaguely defined, in the systems biology community, when connected with the informativity of actual experimental data. In addition, simplistic views of the laws relating measurements with the observed phenomena, as established *e.g.* for experiments at equilibrium, may loose their viability when confronted with complex experimental conditions, such as transient dynamics, and need to be reconsidered with great mathematical care [32].

In this chapter we discuss quantitative reconstruction of regulatory interactions in two cases of practical interest. We first focus on reconstruction of quantitative, dynamical gene regulatory network models from time-course fluorescent reporter gene expression data. Then, we move on to the problem of inferring regulatory laws of metabolism from steady-state metabolite concentration and reaction rate (flux) measurements. In both cases, the publications we concentrate on represent the finalization of work started in previous papers where reconstruction methods are developed. Here we dig deeper into the analysis of the reconstruction problem and the challenges of applying reconstruction methods to real data.

In the first case, we apply the so-called sign-pattern method previously developed in [89] to the reconstruction of the gene regulatory interactions of a well understood *E.coli* motility module. The results, fully discussed in [107], show in particular the importance of recovering the activity profile of a promoter, the physically regulated entity, from the dynamically related time-course fluorescent reporter abundance, the measured quantity, for successful network reconstruction. In fact, this dynamical inversion problem is an example of the general problem of recovering the biologically relevant information from indirect data by dedicated modelling and analysis. Another instance of this problem is treated in a single-cell context in Chapter 3.

In the second case, we start from the so-called lin-log metabolic regulation modelling approach, used in [9] to develop a reconstruction method in presence of incomplete datasets, to pursue a unified investigation of identifiability from a theoretical and practical viewpoint. These two notions refer respectively to the idealized case of fully informative data, and to the presence of finite noisy datasets. Fully developed in [10], the work also includes methods for model reduction and an application to a state-of-the-art dataset from the *E.coli* central carbon metabolism, which illustrates well the pertinence of these questions.

1.2 Reconstruction of gene regulatory networks

In essential terms, gene expression is the biochemical process leading to the synthesis of new molecules of the protein encoded by the corresponding gene. Expression occurs upon activation of the corresponding promoter, which can be regulated by the products of other genes, thus giving rise to a network of interactions. Mathematically, this can be represented by a graph with n nodes $\mathcal{N} = \{i : 1, \ldots, n\}$ for genes and (directed) edges for (directed) interactions. Associating a protein abundance x_i to every gene i, regulated

synthesis of x_i can be quantified by a function $g_i(x, \theta)$, where $x = (x_1, \ldots, x_n)$, and θ are typically unknown parameters, which we omit from notation where convenient for simplicity (direct dependence on time is also possible *e.g.* in presence of environmental perturbations, but we will not discuss this here).

Quantitative regulatory network reconstruction is the problem of inferring g_i from data, with $i \in \mathcal{N}$, from within a suitable class of functions $\mathscr{G}$. Let us initially assume that, for every i, perfect measurements $x^k = (x_1^k, \ldots, x_n^k)$ of x and corresponding values g_i^k of $g_i(x^k, \theta)$ are available for some set of indices k (corresponding *e.g.* to different measurement instants in a time-course experiment). For any given i, we call $\mathscr{D}_i$ the dataset composed of the couples $\{(x^k, g_i^k),\ \forall k\}$. In a time-course experimental setup, these values may be (approximately) recovered from time-lapse quantification of gene expression. The precise relation between experimental measurements and $\mathscr{D}_i$ is a point of utter importance that we will come back to in a moment.

Reconstruction subsumes the identification of the interaction graph, that is, for every i, the effective dependencies of g_i on the elements of x. For well-structured regulatory function families $\mathscr{G}$, knowledge of the interaction graph can greatly simplify the identification of the regulatory functions $g_i \in \mathscr{G}$ and of their parameters θ. A method to reconstruct the interaction graph from the datasets $\mathscr{D}_i$ was proposed in [89], together with an algorithm to identify the regulatory functions and parameters, and their generalization to noisy data. Here, for the later application in [107], we restrict attention to the method for interaction graph reconstruction.

Following [89], we assume that the elements of $\mathscr{G}$ are monotonic functions of every element of x. This is justified in view of the typically monotonic effects of regulators on every given target gene, and also in view of observability considerations [45]. Focusing on any given regulation target gene i, one can qualitatively represent regulatory effects by discrete variables $p_{i,j} \in \{-1, 0, +1\}$, with $j \in \mathcal{N}$, defined such that $p_{i,j} \cdot g_i$ is a nondecreasing function of x_j, with $p_{i,j} = 0$ in the special case where g_i is independent of x_j. We call $p_i = (p_{i,1}, \ldots, p_{i,n})$ the sign pattern of g_i. Sign patterns correspond to directed, signed arcs of an interaction graph (in particular, p_i corresponds to the incoming arcs of node i; see Fig. 1.1 for an example).

Assuming p_i is the true sign pattern of g_i, for any two couples (x^k, g_i^k) and (x^l, g_i^l) from $\mathscr{D}_i$, it must hold that

$$p_{i,j} \cdot (x_j^k - x_j^l) \geq 0, \quad \forall j \in \mathcal{N} \implies g_i^k - g_i^l \geq 0. \tag{1.1}$$

A sign pattern that falsifies (1.1) is called inconsistent (with $\mathscr{D}_i$). In practice, one can efficiently construct a set $\bar{P}_i$ of inconsistent sign patterns by exploration of $\mathscr{D}_i$. For every index pair (k, l) such that $g_i^k - g_i^l < 0$, it suffices to define a new element $\bar{p}_i$ of $\bar{P}_i$ by setting $\bar{p}_{i,j} = \mathrm{sign}(x_j^k - x_j^l)$, $j \in \mathcal{N}$. It is shown in [89] that any inconsistent pattern can be deduced from an element of $\bar{P}_i$ by turning some of its nonzero entries to 0, and that a set of so-called minimal consistent sign patters P_i^* is easily computed from $\bar{P}_i$, such that any pattern verifying (1.1) can be deduced from an element of P_i^* by turning some zero entries to 1 or -1. This hierarchical structure gives rise to efficient algorithms that we do not discuss here, see [89].

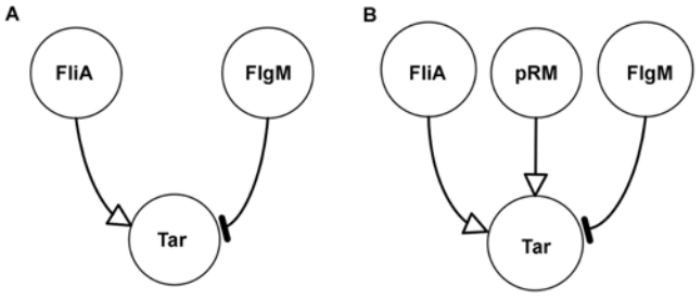

Figure 1.1: Examples of gene regulatory patterns from [107]. Sign patters correspond to directed graphs with positive (arrow-end) or negative (butt-end) regulatory effect on the target gene (zero entries of a sign pattern correspond to absence of an arc). (A) True and correctly inferred minimal sign pattern for the *E.coli* motility module studied in [107]. Other patterns consistent with data are formed by adding regulatory arrows. (B) Same, with the additional regulatory effect of overall cell physiology (monitored via the pRM gene, see paper for details).

In order to study the network reconstruction problem on real time-lapse data, in [107], we applied sign pattern analysis to gene expression measurements from the *E.coli* motility module composed of genes Tar, FliA and FlgM, a well-understood module that provided us with a ground-truth to benchmark results (see Fig. 1.1). Because sign pattern analysis is based on minimal assumptions, results are indicative of the intrinsic complexity and issues of network reconstruction.

We focused on gene expression measurements obtained by the use of fluorescent reporter systems (see Fig. 1.2). This is a synthetic gene construct ensuring that fluorescent protein molecules, say F, are synthesized alongside the protein molecules of the gene of interest, say P, thus providing a way to monitor expression activity over time and *in vivo*. In a deterministic dynamical setup, in the simplest case, expression of F alongside P for the ith gene can be modelled as [30]

$$\begin{aligned} \dot{x}_i &= -\gamma_P x_i + g_i(x), \\ \dot{y}_i &= -\gamma_F x_i + g_i(x), \end{aligned} \tag{1.2}$$

where, in accordance with previous notation, x_i is the concentration of P, g_i is the regulation function of the promoter activity, and the new variable y_i quantifies the concentration of F. Parameters γ_P and γ_F are degradation constants that are generally different (in [107] a slightly more complex model was used, but the main points can be discussed on the basis of (1.2)).

At every measurement time t, for the various genes i, the measured quantity is $y_i^k = y_i(t_k)$. In traditional steady-state experiments, x_i and y_i are both proportional to g_i, thus all these quantities can generally be interchanged. In dynamical conditions, instead, $y_i(t)$ is by no means proportional to $x_i(t)$ or $g_i\big(x(t)\big)$. Unfortunately, the legitimate steady-state assumption that any of these quantities equally represents gene expression strength is often carried over to dynamical scenarios, where a necessary condition for its correctness is that $\gamma_P = \gamma_F$.

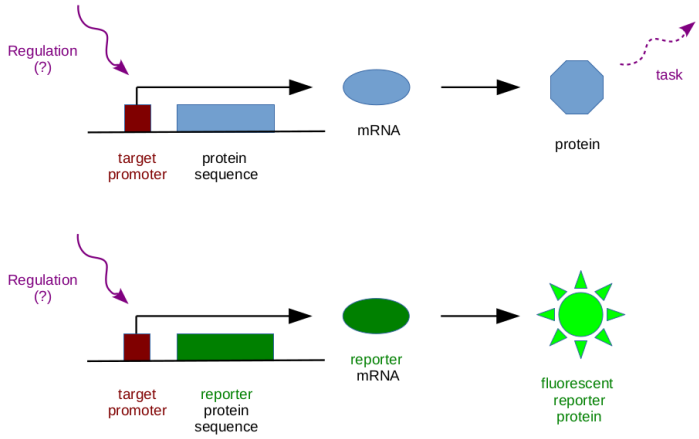

Figure 1.2: Illustration of gene expression and fluorescent protein reporting. (Top) gene expression involves transcription of mRNA molecules and their translation into new molecules of the protein species encoded by the gene (all molecules are furthermore subject to degradation). Trascriptional activation depends on regulation of the gene promoter by molecular species referred to as transcription factors. Synthesized proteins may exert further regulatory effects in the cell, *e.g.* they can themlselves be transcription factors for the expression of other genes. (Bottom) Fluorescent reporter molecules are synthesized in response to the activation of the same promoter via usually different transcription and translation dynamics.

In [107], we studied the impact of erroneous hypotheses on the data, as well as the role of often neglected regulatory factors, on network reconstruction. Based on our own experimental data, we first analyzed the reconstructed sign patterns under the common but wrong estimate that $x_i^k \simeq g_i^k \simeq y_i^k$, for all $i \in \mathcal{N}$. Then, provided relevant values for γ_P and γ_F, we repeated the sign pattern analysis on estimates of (x^k, g_i^k) drawn from a transformation of the measurements y_i explicitly accounting for the dynamical relationships (1.2). To complete the quantitative analysis, we also used a biologically relevant model of $g(x)$ to predict the activation profile of the Tar promoter on the basis of the two different estimates of the x profiles. We additionally considered the role of overall cell physiology as a source of aspecific regulation (captured by an additional regulatory variable x_i along with its own experimental measurements).

The results of the work showed the importance of utilizing an appropriate measurement model for both network reconstruction and prediction of regulated promoter activity profiles. Lack of an appropriate data model as *e.g.* (1.2) was indeed shown to significantly bias network reconstruction. Results also shed light on the importance of global cell physiology, which was found to be an essential regulatory factor for correct quantitative model predictions. This fact is not entirely acknowledge in the biology community (but see [11]) and is a biological contribution of the work. From the methodological viewpoint, besides reconfirming the interest of sign pattern analysis, this work provides the tools and guidelines for sound gene network reconstruction from fluorescent reporter data. The work also bears evidence of how sound methodological

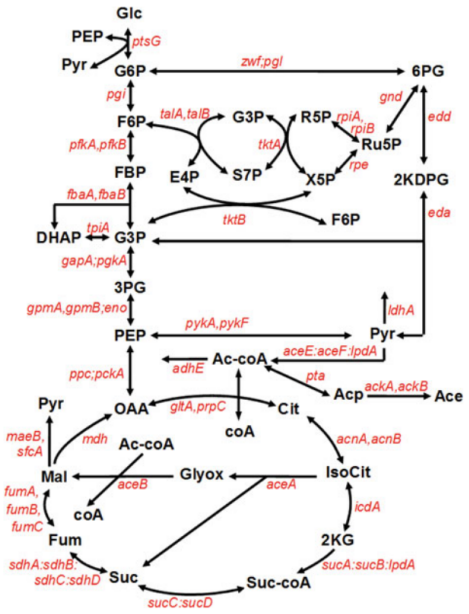

Figure 1.3: Reaction diagram of the *E.coli* central carbon metabolism (see [10]). Metabolites are shown in bold, arrows show metabolite conversion reactions, genes producing the reaction-mediating enzymes are in italic (red).

approaches can help biological discovery. Despite the relative robustness to inaccuracies in model (1.2), the work points at the importance of accurate estimation of the hidden, biologically relevant quantities from the experimental measurements, and stimulates further research in this domain. Results are the fruit of the Ph.D. work of Diana Stefan, centered on the application of my previously started work on an experimental case study, which I co-supervised most notably on the aspects of sign pattern analysis, mathematical modelling and data processing.

1.3 Reconstruction of metabolic regulation models

At a cellular level, metabolism refers to the ensemble of biochemical interactions devoted to transforming substrates into energy and molecular products for the growth and maintenance of the cell (see Fig. 1.3 for an example). Conversion of substrates is generally mediated by enzymes, that is, products of gene expression, and inversely, metabolite products can participate in gene regulation. It is thus clear that a detailed mathematical description of metabolic reaction dynamics is daunting.

In the literature, one approach to infer quantitative models of metabolism is by the use of phenomenological models. These are mathematical laws describing how observed reaction velocities depend on other observed quantities (enzymes, metabolites) rather than the mechanisms that yield these laws. Often formulated in terms of a parametric model class, identification of metabolic models then amounts to estimate the values of

the model parameters from experimental data.

Among others, one class of models that is well adapted to describe metabolism regulation models is the lin-log (linear-logarithmic) models. These are models in the form

$$v(x,u,e) = \mathrm{diag}(e) \cdot \big(a + B^x \cdot \log(x) + B^u \cdot \log(u)\big) \tag{1.3}$$

where v is the vector describing reaction rates, x (resp., u) is a vector of internal (resp., external) metabolite concentrations, and e is a vector of reaction-mediating enzyme concentrations. Parameters (a, B^x, B^u) are generally the unknowns to be estimated from experiments. Being the model linear in the log-domain, this is a particular instance of pseudo-linear models. Thanks to the shape of the log function, well mimicking the commonly observed Michaelis-Menten laws around a nonzero reference point, this model structure strikes a very favorable compromise between descriptive power and tractability.

State-of-the-art experimental techniques allow one to measure enzyme levels and corresponding metabolite concentrations in steady-state in several conditions (*e.g.* in response to different environmental conditions u) [53]. Knowing that concentration dynamics obey $\dot{x} = N \cdot v(x,u,e)$, where N is the reaction network stoichiometry matrix, this corresponds to measurements obtained in a state that satisfies $0 = N \cdot v(x,u,e)$. Time-lapse measurements of x and e are instead difficult to obtain and were not the focus of our research.

Typically in this context, the stoichiometry matrix N and the regulatory interactions (nonzero elements of B^x and B^u) are considered known or fixed upon lumping of non-observable reactions. Observations of (e, v, x, u) in different conditions, which are generally noisy, may instead be incomplete, that is, have missing entries due to issues in the measurement process. In [9], under appropriate assumptions on the measurement noise, a method for the estimation of $\theta = (a, B^x, B^u)$ from incomplete datasets was developed based on Expectation-Maximization and a data-driven definition of Bayesian priors for the missing data entries. The method was shown to provide excellent results in simulation as well as on the real data from the *E. coli* central carbon metabolism in [53] (see also Fig. 1.3).

Results of [9] raised questions concerning the identifiability of (1.3), as a function of the choice of the external conditions u and of the noise affecting the data. In the case of complete data, the identification problem is easily reduced to a linear least-squares problem. From this perspective, in absence of noise, uniqueness or equivalence of solutions for θ reduce to algebraic properties such as the structure of certain nullspaces, and is essentially determined by the choice of u. On the other hand, the precision by which the solution (or subset of equivalent solutions) can be determined is a property that also depends on the measurement noise model. In [10], we refer to the first as theoretical identifiability and to the second as practical identifiability.

In the systems biology literature, practical identifiability always turns out being a condition checked a-posteriori on the basis of a specific estimation procedure, essentially aimed at testing the reliability of a specific parameter estimate. In so doing, properties of the model are blurred with properties of the chosen estimator, and are attached to

the fate behind a specific dataset. This hampers the analysis of the information that data could provide about the model if duly exploited, and prevents one from putting theoretical and practical identifiability into a clear relation.

In [10], for lin-log models in particular and pseudo-linear models in general, we reviewed theoretical identifiability in a sound algebraic context, also providing tools to spell out equivalent solutions and extract well-defined parameters of a nonidentifiable model when possible. Then, we addressed practical identifiability in the perspective of a priori analysis and relate it explicitly with theoretical identifiability. Faithful to the very term identifi*ability* denoting a property and not the outcome of an estimation procedure, we formulated practical identifiability as the question of existence of an estimator satisfying prespecified requirements on estimation accuracy, and showed that theoretical identifiability is not only necessary for practical identifiability, but also sufficient to guarantee non-degenerate estimates. It is worth emphasizing that the work does not boil down to a linguistic exercise around the term identifiability, since it has the practical value of addressing the question 'What can be obtained?' rather than 'What have we obtained?' from a given model and experiment.

Of course, it can be argued that this analysis is made possible by the favorable structure of the model class. Still, important concepts are put forward that in my opinion should be pursued in all variants of the problem (via simulation if analytical treatment is impossible). First, practical identifiability is a concept that cannot be disentangled from requirements on estimation accuracy. Existing definitions such as boundedness of confidence intervals [99] seem unsatisfactory, since arbitrarily large confidence intervals are as unacceptable as unbounded ones. Second, practical identifiability should be a joint property of the model and the experiment, not of a specific estimation procedure or dataset. Otherwise, one may deem unexploitable an informative dataset, or inappropriate a valuable experiment design. Third, lack of theoretical identifiability should not be equated to model uselessness. A characterization of all equivalent solutions compatible with data is possible and often very informative.

The methods and concepts developed in [10] were also illustrated by means of several examples and applied on the same real state-of-art dataset used in [10], also exploiting the methods for incomplete datasets discussed in the same paper. This case study served us to show what serious lack of (practical) identifiability affects these problems (a summary of the identifiability analysis results is in Fig. 1.4), and yet what biologically relevant conclusions can be drawn *e.g.* concerning dependencies among metabolites or reactions operating close to equilibrium. The study was the result of the work with the Ph.D. student Sara Bertoumieux, whom I started to co-supervise at my arrival in the Inria team IBIS [51], contributing much of the mathematical and statistical pitch of the research.

	Enzyme \ Metabolite	Glc	PEP	G6P	Pyr	F6P	FBP	DHAP	3PG	Ac-coA coA	6PG	Ru5P	R5P	S7P	2KG	Suc	Fum	Mal	ATP ADP	Cit	NADPH NADP	NADH NAD	FAD	Ace
1	PtsG	100%	100%	31%	0 %																			
2	Pgi			0 %		0 %																		
3	PfkA,PfkB		70%			2%	63%												8%					
4	FbaA,FbaB						100%	100%																
5	TpiA							100%	100%															
6	GapA;Pgk						84%		6%										0%			51%		
7	GpmA,GpmB;Eno		0%						0%															
8	PykA,PykF		5%		0%		22%												0%					
9	AceE:AceF:LpdA				1%					99%												94%		
10	Zwf;Pgl			70%							100%										2%			
11	Gnd										100%	100%									0%			
12	Rpe											0%		0%										
13	RpiA,RpiB			0%								0%	66%											
14	TktA												82%	82%										
15	TalA,TalB					0%								0%										
16	TktB					1%						1%												
17	GltA,PrpC									97%					2%					98%		97%		
18	AcnA,AcnB														0%					100%				
19	IcdA									100%					0%						0%			
20	SucA:SucB:LpdA;SucC:SucD														0%	0%						90%		
21	SdhA:SdhB:SdhC:SdhD															0%	0%						62%	
22	FumA,FumB,FumC																0%	0%						
23	Mdh		87%															0%		75%		35%		
24	Ppc;PckA		19%				5%											0%	0%	32%				
25	MaeB,SfcA				0%					99%								0%			0%	30%		
26	AceA;AceB									100%						0%		0%						
27	μ		0%	0%	0%	0%			0%	68%			81%		0%				0%		0%	27%		
28	Edd:Eda				0%						100%													
29	Pta;AckA,AckB				0%					70%									0%		0%	37%		100%
30	LdhA				0%																	95%		
31	AdhE									100%														
	% Missing data	3	17	0	48	7	34	59	10	3	72	3	38	3	59	3	14	14	0	62	79	79	17	17

Figure 1.4: Summary of *E.coli* central carbon metabolism identifiability results from the analysis of a state-of-the-art dataset in [10]. Identifiable regulation parameters are highlighted in green. Reactions in green have all parameters identifiable. Nonidentiable parameters are in grey. Percentages report identification success rate relative to repeated imputation of missing data entries (see [10] for details on the analysis).

1.4 Discussion and perspectives

In this chapter I presented an overview of my research activity concerning reconstruction of quantitative models of gene expression and metabolic networks from experimental biology data. Starting from earlier work where reconstruction methods were developed, focus was placed on several conceptual and practical problems that quantitative network reconstruction may conceal at a first sight. In a deterministic modelling context, naturally matching population-average data, much has been done in the literature, although space for innovation is still left especially in a dynamical context. Yet, most fundamental issues investigated in this chapter, *e.g.* identifiability and model reduction, are still of great interest and actively investigated, partly due to the difficulty of finding analytic solutions in a general nonlinear context. Moreover, network reconstruction problems present themselves in different forms depending on the new types of available data. Along these lines, my own first efforts toward stochastic network reconstruction from population snapshot data are the object of [21] (see also Section 3.3).

Two general lessons can be learned from this work and relevant work in the literature. First, most often, model reconstruction problems should not aim at reconstructing a unique model, but rather a family of models compatible with the experimental evidence. Whether in the form of alternative network structures or of indistinguishable model parametrizations, this applies in both examples analysed, and beyond. Depending on the context, nontrivial solution sets can be sufficient for the modelling goals, or may call for further refinements to be sought by more specific analysis and experiments. If, on the one hand, this points at the limitations of general formulations of network reconstruction problem, on the other hand, it opens up intriguing and profitable re-

search questions in the context of so-called ensemble modelling (characterizing sets of compatible models [60], similar in spirit from the equivalent parameter sets mentioned in Section 1.3) as well as experimental design. In fact, for the case of single-cell models, questions revolving around optimal experiment design are part of the ANR project MEMIP that I participate in [75].

The second lesson is the importance of appropriate data analysis. As it has been shown most notably in the gene network example, this goes well beyond simple data processing protocols. Depending on the data to be treated and the questions to be addressed, a whole set of challenges arise that require dedicated expertise at the intersection of mathematical modelling, statistics, theoretical and experimental biology. Some questions of this kind coming from current experimental techniques are being addressed here in Chapter 3. In perspective, there is no doubt that new upcoming experimental techniques will necessitate equally innovative theoretical and methodological contributions for the correct and in-depth exploitation of the data. Concrete examples of this need from my recent and near-future research can also be recognized in Chapter 2.

[107] Inference of quantitative models of bacterial promoters from time-series reporter gene data (2015). D.Stefan, C.Pinel, S.Pinhal, E.Cinquemani, J.Geiselmann, H.de Jong), *PLoS Computational Biology*, 11(1):e1004028.

Abstract: The inference of regulatory interactions and quantitative models of gene regulation from time-series transcriptomics data has been extensively studied and applied to a range of problems in drug discovery, cancer research, and biotechnology. The application of existing methods is commonly based on implicit assumptions on the biological processes under study. First, the measurements of mRNA abundance obtained in transcriptomics experiments are taken to be representative of protein concentrations. Second, the observed changes in gene expression are assumed to be solely due to transcription factors and other specific regulators, while changes in the activity of the gene expression machinery and other global physiological effects are neglected. While convenient in practice, these assumptions are often not valid and bias the reverse engineering process. Here we systematically investigate, using a combination of models and experiments, the importance of this bias and possible corrections. We measure in real time and in vivo the activity of genes involved in the FliA-FlgM module of the *E. coli* motility network. From these data, we estimate protein concentrations and global physiological effects by means of kinetic models of gene expression. Our results indicate that correcting for the bias of commonly-made assumptions improves the quality of the models inferred from the data. Moreover, we show by simulation that these improvements are expected to be even stronger for systems in which protein concentrations have longer half-lives and the activity of the gene expression machinery varies more strongly across conditions than in the FliA-FlgM module. The approach proposed in this study is broadly applicable when using time-series transcriptome data to learn about the structure and dynamics of regulatory networks. In the case of the FliA-FlgM module, our results demonstrate the importance of global physiological effects and the active regulation of FliA and FlgM half-lives for the dynamics of FliA-dependent promoters.

[10] On the identifiability of metabolic network models (2013). S.Berthoumieux, M.Brilli, D.Kahn, H.de Jong, E.Cinquemani, *Journal of Mathematical Biology*, 67(6-7):1795–1832.

Abstract: A major problem for the identification of metabolic network models is parameter identifiability, that is, the possibility to unambiguously infer the parameter values from the data. Identifiability problems may be due to the structure of the model, in particular implicit dependencies between the parameters, or to limitations in the quantity and quality of the available data. We address the detection and resolution of identifiability problems for a class of pseudo-linear models of metabolism, so-called linlog models. Linlog models have the advantage that parameter estimation reduces to linear or orthogonal regression, which facilitates the analysis of identifiability. We develop precise definitions of structural and practical identifiability, and clarify the fundamental relations between these concepts. In addition, we use singular value decomposition to detect identifiability problems and reduce the model to an identifiable approximation by a principal component analysis approach. The criterion is adapted to real data, which are frequently scarce, incomplete, and noisy. The test of the criterion on a model with simulated data shows that it is capable of correctly identifying the principal

components of the data vector. The application to a state-of-the-art dataset on central carbon metabolism in *Escherichia coli* yields the surprising result that only 4 out of 31 reactions, and 37 out of 100 parameters, are identifiable. This underlines the practical importance of identifiability analysis and model reduction in the modeling of large-scale metabolic networks. Although our approach has been developed in the context of linlog models, it carries over to other pseudo-linear models, such as generalized mass-action (power-law) models. Moreover, it provides useful hints for the identifiability analysis of more general classes of nonlinear models of metabolism.

Chapter 2

Identification of stochastic single-cell gene expression models

2.1 Intercellular variability of gene expression dynamics

Modern experimental techniques allow one to monitor gene expression response dynamics to environmental stimuli in individual microbial cells [110]. Quantitative time-course gene expression measurements, such as obtained by the combined use of fluorescent reporter proteins, microfluidic devices, and videomicroscopy, reveal that, even in isogenic cell populations, expression dynamics are highly variable. Variability takes the form of fluctuations in the single-cell response as well as quantitative differences across different cells [37]. It is believed to be at the roots of crucial population-level phenomena such as bet-hedging and diversification, as well as of the emergence of specific intracellular regulatory patterns to counteract expression uncertainty where this could be detrimental [95].

Toward a quantitative investigation of origin and consequences of gene expression variability, a significant effort has been devoted to its mathematical modelling in the recent years. Several modelling approches have been proposed to describe intrinsic noise (that is, the randomness inherent in the gene expression transcription and translation reactions), extrinsic noise (that is, the additional uncertainty affecting gene expression dynamics via regulatory factors and overall cell physiology), or both [43, 87, 109, 50, 62, 112]. Correspondingly, attention has been paid to the problem of fitting the theoretical models to the available single-cell data. Despite diverse contributions [129, 128, 81, 79, 41, 57], the problem to find an appropriate modelling approach combining descriptive power and ease of reconstruction from experimental data is far from being settled.

The choice of the modelling approach and of the inference procedure depends in the first place on the specific biological system, experimental condition and data, but also on the scientific question being addressed. Of course, a model capturing all possible

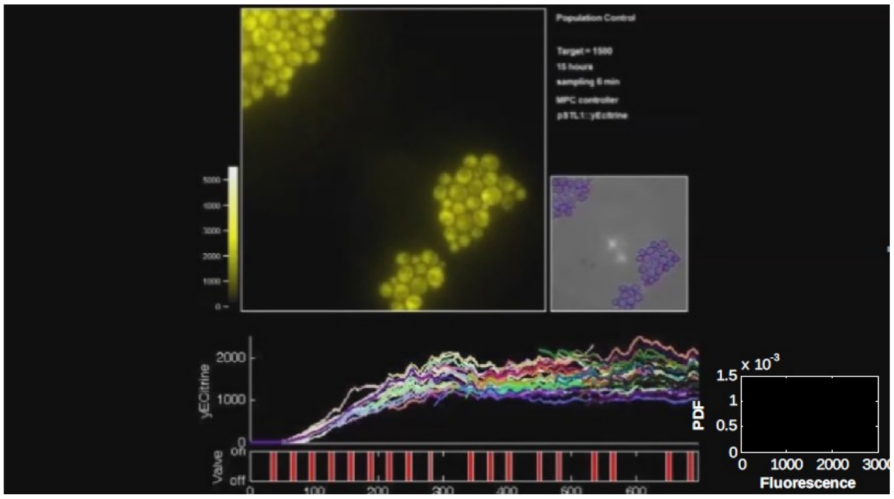

Figure 2.1: Example of fluorescence videomicroscopy for the monitoring of yeast osmotic shock gene expression response variability in single cells [67]. (Top Left) Fluorescence onset in cells grown in microfluidic devices and subjected to osmotic shocks. (Top right) Cell tracking by image processing. (Bottom) Computed fluorescence levels in single cells over time (individual traces) in response to repeated shocks (red bars indicate instants and duration of shocks).

sources of variability in detail is out of question for complexity, insufficient knowledge of the biochemical processes at play, and limited informativity of the data. In the sought of a convenient compromise, work devoted to intrinsic noise has been revolving around stochastic dynamical models such as continuous-time Markov chains and continuous-state approximations thereof, accompanied by tools based on the Chemical Master Equation, Moment Equations, and Gillespie simulation for their inference from data [119, 120]. However, renewed interest has been directed to extrinsic noise, motivated by the fact that in many cases this appears to be the dominant source of variability [97, 26, 88].

Different from intrinsic noise, extrinsic noise can often be associated with fluctuations occurring at a time scale comparable with that of the cell cycle [97, 37]. Thus, in a first reasonable approximation, one may picture gene expression dynamics as invariant in every cell but different across cells. Provided intrinsic noise is negligible (as *e.g.* for gene expression processes characterized by high copy numbers of mRNA and protein molecules [42]), for the investigation of variable individual cell identity in a homogeneous population, this hints at the utilization of deterministic expression models with parameters that take different values for different cells. This is the object of the modelling approach described in Section 2.2, which summarizes the contributions of [67], where each individual cell is treated as statistically independent from the others. Such independence is however an oversimplification especially when couples of mother and daughter cells can be observed. Indeed, because daughter cells inherit their material from the mother, one may expect that intercellular variability is more pronouced for cells that are far apart in the population lineage. For the case where parental re-

lationships among the experimentally observed cells are available, an extension of the modelling and inference approach of Section 2.2 is the object of Section 2.3, summarizing the contribution of [72].

2.2 ME modelling and inference from single-cell data

In [67], we considered modelling of intercellular gene expression variability in yeast cells in response to osmotic shocks. We focused on individual-cell time-course gene expression reporter data for cells grown in microfluidic devices and observed by videomicroscopy, with a fluorescent reporter protein placed under control of the osmoresponsive gene STL1. Suitable experiment design allowed us to assume that the observed dynamics in response to time-varying osmotic perturbations are not altered by unwanted regulatory or adaptation effects.

In order to model intercellular variability of the gene expression dynamics, we considered a deterministic response model with cell-dependent parameters. The system of equations

$$\begin{aligned} \dot{m}(t) &= -g_m m(t) + k_m u(t), \\ \dot{p}(t) &= -g_p p(t) + k_p m(t), \end{aligned} \tag{2.1}$$

with $m(t)$ and $p(t)$ the cellular concentration at time t of the mRNA and protein molecules synthesized from the gene, and $u(t)$ the level of activity of the gene, is a commonly accepted description of the gene expression transcription (mRNA synthesis) and translation (protein synthesis) dynamics. For the data at hand, $u(t)$ is the (known) promoter activity in response to the osmotic shocks, and $p(t)$ is the (measured) fluorescence level (additional modelling details concerning *e.g.* protein maturation and the relation between u and osmotic shock profile can be found in [67]). Kinetic parameters $\psi = (k_m, g_m, k_p, g_p)$ take a different value for every cell. Thus, the identity of a cell in a population of a priori identical cells is defined by its own parameter values ψ^v, with v denoting the generic vth cell from within a set of cell indices V.

A possible approach to reconstruct such model from the data is to fit the parameters of every cell to the corresponding gene expression profiles, possibly studying the statistical distribution of the paramenters over the population in a second step. This approach, which we referred to in [67] as the naive approach, is bound to fail in presence of a small number of noisy measurements per cell. In addition, it fails to account for the fact that these parameters describe cells of a same homogeneous population. The approach we proposed relies on ME modelling [61]. Introduced in the context of pharmacokinetics to model response of different individuals to drug delivery, this approach postulates that the individual parameters are statistically independent random outcomes from a common population distribution,

$$\psi^v \sim \mathscr{D}(\theta) \tag{2.2}$$

where the parameters θ of the probability distribution $\mathscr{D}$ characterize the population. In particular, fixed population effects as well as random individual effects can be incorporated into this definition, whence the name of the approach (more extensive formulations

of ME modelling can be found *e.g.* in [61]). In practice, for positive parameters, a common choice is the log-normal distribution. That is, for $\varphi^v = \log \psi^v$, one postulates the normal distribution

$$\varphi^v \sim \mathcal{N}(\mu, \Sigma) \tag{2.3}$$

where $\theta = (\mu, \Sigma)$ are the population parameters. Rather than processing every cell separately, ME identification methods estimate θ first from the whole set of measurements from all individuals, while treating individual parameter estimates a posteriori. Thus, the fact that individuals are part of a homogeneous population is inherent in definition (2.2). In addition, usage of this relation in the identification process in known to improve the estimation of both population parameters θ and individual parameters ψ^v.

The application of the ME paradigm to yeast data in [67] (with the lognormal hypothesis (2.3)) and its comparison with the naive approach reconfirmed the expectations. Despite a relative abundance of individual cell measurements for certain cells (due to asymmetric division of budding yeast, cells giving birth to new daughters were considered to be the same cell before and after division), ME provided a dramatic improvement in the estimation of population parameters. Given the absence of ground truth, this was verified by comparing model predictions (statistical distribution of protein levels at different times) from ME identification and naive identification with data from new perturbation experiments on cells with identical genome. The difference between the accuracy of ME-based and naive-based predictions was enormous. ME also detected stronger correlations among the entries of the individual-cell parameter vector ψ, which was found to be an important factor for the increased predictive capability. All this can be understood in terms of the use of assumption (2.2), which introduces a structure in the reconstruction of the population distribution of individual parameters.

From the biological viewpoint, statistical analysis of the ME a-posteriori individual cell estimates in relation with other recorded single-cell properties (size, position in the colony, etc.) not used in identification revealed interesting correlations. Most notably, leveraging known parental relationships among the observed cells, analysis of individual-cell parameter estimates revealed a statistically significant correlation between mother and corresponding daughter cell parameters. This observation agrees with intuition, since the biochemical material that determines cellular regulatory dynamics is transmitted at cell division, and deserves further investigation. At the same time, it falsifies the typical ME modelling assumption of statistical independence across different individual parameters ψ^v, which makes standard ME modelling inappropriate for this investigation. These observations are at the roots of the work described in the next section.

Our work [67] represented the first application of the ME paradigm to single-cell gene expression time-course data. Despite the conceptual relevance, practical application to real data is nontrivial in several respects, such as the characterization of measurement noise, the evaluation of the identified model quality and (as already discussed in Chapter 1) structural and practical identifiability issues (let apart experimental and raw data processing aspects). The work has been developed in collaboration with the Inria/Pasteur group InBio [52], with the MSC laboratory of the Université Paris-Diderot [78], and with the University of Pavia in Italy (Giancarlo Ferrari-Trecate). It

involved the activity of the Ph.D. student Andres Gonzales-Vargas, whom I supervised in his 6-month visit in IBIS from the University of Pavia. From a purely technical viewpoint, my main contributions were in the choice of the individual response model to ensure well-posedness of the identification problem (much more complicated models were considered and discarded due to severe lack of identifiability), the definition of a structurally identifiable parametrization, and the investigation of the role of the population model in the identifiability of individual-cell parameters (see the supplementary material of the paper).

2.3 ARME modelling and inference from single-cell and lineage tree data

Following up from the results described above, in [72], we addressed the question of gene expression kinetics inheritance from mother to daughter cells. In the light of the observed correlation between mother and daughter cell parameters, the main methodological question addressed in [72] is whether accounting for possible parameter inheritance at the very modelling stage improves identification results in some quantifiable sense. From the biological viewpoint, the assessment of the degree of inheritance of gene expression parameters may shed new light on how the observed variability across cells is built up along generations, with potential implications in the study of phenotypic diversity in the face of changing environments.

Concretely, we proposed a generalization of the population model (2.3). Instead of assuming statistical independence across individuals, we assume that the parameters of mother cell v^- and daughter cell v obey the stationary Auto-Regressive (AR) model (see also Fig. 2.2 below)

$$\varphi^v = A\varphi^{v^-} + (I - A)b + \eta^v \tag{2.4}$$

with $\eta^v \sim \mathcal{N}(0, \Omega)$ independent across v. By the stationary assumption, for all $v \in V$, parameters φ^v are distributed as in (2.3) with $\mu = b$ and Σ obeying the Ljapunov equation $\Sigma = A\Sigma A^T + \Omega$. However, they are no longer independent. In particular, matrix A admits the interpretation of (matrix) correlation coefficient between mother and daughter parameters. The model naturally accommodates two (or more) daughter cells, as in symmetric division, as well as budding yeast asymmetric division, which is the case of [67]. In the latter case, the mother cell is reasonably assumed to keep its own parameters after division (more complex scenarios can be considered by suitable extensions of (2.4)). Together with (2.1), we call (2.4) the ARME modelling of gene expression. This constitutes the first contribution of [72]. Importantly, it includes ME modelling of gene expression as a special case with $A = 0$.

In ARME modelling, the parameter triplet $\theta = (A, b, \Omega)$ defines all joint population statistics. In terms of identification, in presence of lineage information $W \subseteq V \times V$ (a set of mother-daughter cell pairs), the objective becomes that of estimating θ from the whole set of experimentally measured individual-cell time profiles of p, denoted by Y (here we

neglect details and parameters pertaining the measurement model for simplicity, refer to [72] for this). Similar to standard ME identification, estimation of θ can be posed as the Maximum-Likelihood problem

$$\max_{\theta} f(Y|\theta, W), \tag{2.5}$$

where f denotes the relevant probability density function. Different from standard ME identification, parental relations W figure in the problem and are assumed known. We refer to this as ARME identification. To solve the problem, lack of independence across cells does not allow one to use the tools existing for ME identification. A second contribution of [72] has thus been to develop and implement a generalization of the so-called SAEM algorithm [34, 103], a randomized ME identification approach, for the solution of (2.5) (see details and performance assessment in [72]). If of interest, from the estimates of θ, estimates of Σ as well as of individual cell parameters are easily derived from the same algorithm.

For the purpose of assessing inheritance and variability of kinetic parameters at cell division, we classify ARME identification as a direct method. This is in contrast with indirect approaches where the inheritance (correlation) factor A is estimated based on post-processing (empirical correlation analysis) of individual-cell parameter estimates (see Fig. 2.2). In order to answer the methodological question motivating this work, we assessed estimation of parameters θ on simulated data generated with A (diagonal and) strictly positive, and compared the results with the state-of-the-art indirect approach of [67], that is, using individual-cell parameter estimates drawn from the ME approach to build empirical estimates of A. Our ARME direct method was found to be unbiased, while the indirect method systematically underestimated the inheritance (correlation) factor A. A very reasonable explanation of this is that the independence assumption of standard ME implicitly introduces a bias toward a model with $A = 0$, corresponding to independent individuals.

Application to the same yeast osmotic shock response data of [67] was in qualitative agreement with the *in silico* experiments. Compared with the ARME approach, estimates of A provided by the indirect method suggested a much smaller degree of inheritance of gene expression parameters from mother to daughter cells. On the other hand, the two methods returned similar estimates for μ and Σ. Taken together and in view of the unbiasedness of ARME identification, we concluded that our direct approach enables a better assessment of how variability of individual-cell parameters in an isogenic cell population is built up along generations. Moreover, a mother-daugther correlation of about $60\% - 70\%$ irrespective of the specific entry of φ hints that daughter cell parameters are determined by the mother to a fairly large extent, most likely due to the transmission of aspecific regulatory factors. This conclusion opens up interesting biological investigation avenues.

The work described was carried out in collaboration with Aline Marguet in her Post-Doc stay at IBIS. I was the proposer and coordinator of the project, in the broader context of the ongoing ANR project MEMIP where I am principal investigator [75]. The work falls at the forefront of current research in modelling gene expression dynam-

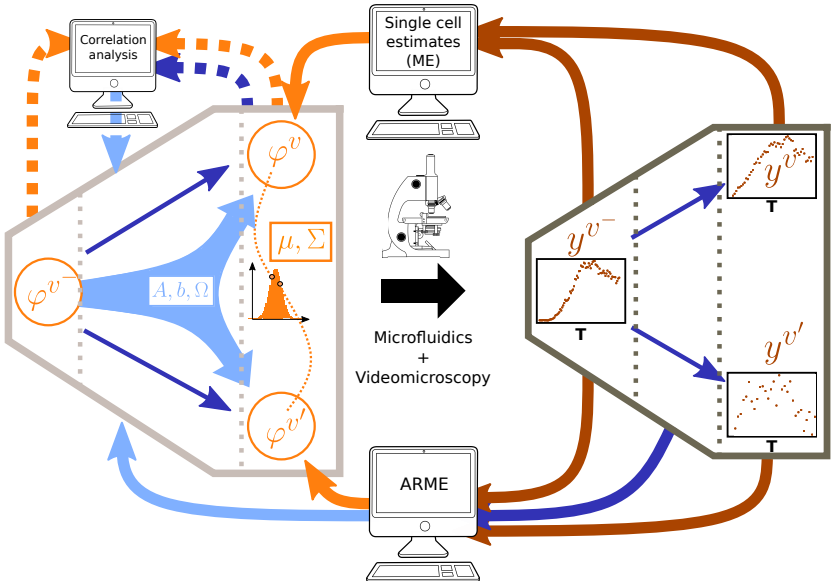

Figure 2.2: ARME versus indirect approaches for the estimation of inheritance and variability of gene expression parameters. (Left) Modelling of single cell parameters as well as their variability and inheritance across cell division. (Right) Experimental measurement of gene expression profiles in the same single cells. Orange circles represent cells, straight blue arrows represent the known parental relationships among them (lineage data). The inference problem considered in this paper is to reconstruct variability and inheritance dynamics (cyan double-arrow, left) of the single-cell parameters (φ^{v^-}, φ^v and $\varphi^{v'}$, orange, left) from gene expression data (y^{v^-}, y^v and $y^{v'}$, red dots, right) and the known parental relationships. Data processing flow from right to left represents utilization of single-cell data (red arrows) and lineage information (blue arrows) to produce estimates of individual cell parameters and statistics (orange arrows) as well as of their variability and inheritance dynamics at cell division (cyan arrows). ARME (bottom) is a direct method that, based on explicit modelling of variability and inheritance dynamics, uses single-cell data together with lineage information to estimate the variability and inheritance parameters (A, b and Ω) at once. Estimates of single-cell parameters and of their statistics (μ and Σ ; orange, left) are also obtained as a byproduct. Indirect (e.g. ME-based) methods (top), instead, only use individual cell data to provide estimates of individual cell parameters and their statistics in a first step. Based on the individual-cell parameter estimates from the first step and lineage information, estimates of inheritance dynamics are produced in a second step.

ics from single-cell data. Despite an existing body of work on the modelling of noise dynamics in populations of dividing cells, this is among the very few works inferring models of expression dynamics from lineage tree data, and the first specifically addressing inheritance and variability of gene transcription and translation parameters at cell division, from both the modelling and the inference viewpoint.

2.4 Discussion and perspectives

The work described in this chapter focuses on extrinsic gene expression noise modelling in isogenic cell populations. More specifically, it relies on ME modelling to describe variability in terms of kinetic gene expression parameters taking different values in different cells. Our ARME extension of the ME paradigm further allows one to determine from lineage data how such variability arises along generations.

The AR process describing the evolution of parameters from mother to daughter cells also provides one possible approach to model extrinsic noise dynamics over time. Yet, this model acts at the time scale of generations. One direction of investigation that I am currently pursuing with Aline Marguet is the extension of ARME to describe parameter fluctuations within a cell lifespan. Though challenging, this would bring our modelling approach closer to the experimental definition of extrinsic noise of [37].

A somewhat complementary direction of investigation is the embedding of intrinsic noise modelling in the ARME paradigm. In essence, this amounts to replace (2.1) with a suitable stochastic counterpart. Irrespective of lineage dynamics, intrinsic noise modelling and identification from single-cell data has been object of much interest in the recent years (see *e.g.* [129]). Yet, a practical intrinsic noise modelling and inference approach in presence of lineage information is not yet established. In this respect, the ARME framework is a promising starting point.

From a biological viewpoint, the question whether intrinsic or extrinsic noise is the dominant source of variability is unresolved and it does not seem to have a unique answer, due to dependence on the specific gene expression system. In a way, the study of noise in presence of lineage data adds a third dimension to the question, namely the importance of noise at cell division. The extensions of the ARME framework discussed above will provide new tools to address this challenging question. The methodological research and the applications on experimental data will still be pursued in collaboration with the InBio team [52] and Marc Lavielle of XPOP [124], in the context of the current project MEMIP [75] and in follow-up projects.

Finally, it should be noticed that addressed in this chapter is the quantitative description of variability in the expression of a single gene. Further efforts connecting our work with the notion of gene regulatory networks are needed in order to contribute to the study of the onset of qualitatively different phenotypic traits along generations. Though apparently far-fetched, the ever-increasing richness of single-cell data, together with the tremendous computational power available nowadays, suggest that such mod-

elling questions and related inference methodologies will be of increasing relevance for the future of microbiology. In perspective, this direction of investigation will not only enrich the available toolkit to study single-cell data with lineage information, but it is also expected to further the understanding of origins and consequences of noise in biological systems.

[67] What population reveals about individual cell identity: Single-cell parameter estimation of models of gene expression in yeast (2016). A.Llamosi, A.M.Gonzalez-Vargas, C.Versari, E.Cinquemani, G.Ferrari-Trecate, P.Hersen, G.Batt, *PLoS Computational Biology*, 12(2):e1004706.

Abstract: Significant cell-to-cell heterogeneity is ubiquitously observed in isogenic cell populations. Consequently, parameters of models of intracellular processes, usually fitted to population-averaged data, should rather be fitted to individual cells to obtain a population of models of similar but non-identical individuals. Here, we propose a quantitative modeling framework that attributes specific parameter values to single cells for a standard model of gene expression. We combine high quality single-cell measurements of the response of yeast cells to repeated hyperosmotic shocks and state-of-the-art statistical inference approaches for mixed-effects models to infer multidimensional parameter distributions describing the population, and then derive specific parameters for individual cells. The analysis of single-cell parameters shows that single-cell identity (e.g. gene expression dynamics, cell size, growth rate, mother-daughter relationships) is, at least partially, captured by the parameter values of gene expression models (e.g. rates of transcription, translation and degradation). Our approach shows how to use the rich information contained into longitudinal single-cell data to infer parameters that can faithfully represent single-cell identity.

[72] Inheritance and variability of kinetic gene expression parameters in microbial cells: Modelling and inference from lineage tree data (2019). A.Marguet, M.Lavielle, E.Cinquemani, *Bioinformatics* (Proceedings of 27th ISMB/18th ECCB), to appear.

Abstract: *Motivation.* Modern experimental technologies enable monitoring of gene expression dynamics in individual cells and quantification of its variability in isogenic microbial populations. Among the sources of this variability is the randomness that affects inheritance of gene expression factors at cell division. Known parental relationships among individually observed cells provide invaluable information for the characterization of this extrinsic source of gene expression noise. Despite this fact, most existing methods to infer stochastic gene expression models from single-cell data dedicate little attention to the reconstruction of mother-daughter inheritance dynamics. *Results.* Starting from a transcription and translation model of gene expression, we propose a stochastic model for the evolution of gene expression dynamics in a population of dividing cells. Based on this model, we develop a method for the direct quantification of inheritance and variability of kinetic gene expression parameters from single-cell gene expression and lineage data. We demonstrate that our approach provides unbiased estimates of mother-daughter inheritance parameters, whereas indirect approaches using lineage information only in the post-processing of individual cell parameters underestimate inheritance. Finally, we show on yeast osmotic shock response data that daughter cell parameters are largely determined by the mother, thus confirming the relevance of our method for the correct assessment of the onset of gene expression variability and the study of the transmission of regulatory factors.

Chapter 3

Estimation of gene expression and metabolic activity dynamics

3.1 Estimation of dynamical quantities in microbiology

Traditional microbiology experiments consist for the largest part of qualitative or semi-quantitative experiments (*e.g.* cell plating, western blots, ...). Often, response of a system to changes in environmental conditions is experimentally observed by very few data points (*e.g.* gene expression microarray measurements before and after a perturbation). Processing of data is frequently carried out on a basic statistical ground or by simple heuristic calculus, with minimal intervention of phenomenological models (such as growth rate being the relative change of population size per unit time). The advent of experimental techniques for the time-course monitoring of bacterial dynamics, notably fluorescent reporters for the dynamical quantification of gene expression, has changed the picture. It is nowadays possible to measure not only growth but also gene expression and other physiological quantities with reasonable time resolution, even at the single-cell level [110]. Analysis of this data naturally calls for more refined data processing and modelling methodologies, and, together with other factors, stimulated the interest of the signal processing, system theory and control communities [62, 116].

Analysis and estimation methods from the system and control theory provide a powerful toolkit for the processing of time-course microbiology data [64, 129, 40]. Yet, gaps still remain toward their full exploitation for biological applications. Despite a general theory, most analysis and estimation algorithms have been developed for systems that do not correspond to the typical scenarios in biology. For instance, asymptotic analysis results from identification theory [66] are hardly applicable to typically short time series, and assumptions that are common in black-box modelling for control, such as linear and discrete-time approximations of the system dynamics, are usually not accepted since too distant from the physically relevant quantities. Conversely, from the viewpoint of the biologist, several misconceptions need to be sorted out to allow for a fruitful application

of these tools. From my experience, for instance, these concern the precise role of measurement models to distinguish measurements from observables and unknowns (see also Section 1.2), or, more advanced, the difference between Kalman filtering and smoothing for causal (possible online) and noncausal (batch) data processing [56].

In the next sections, we describe two problems of this sort that are both relevant to current microbiology research, estimation of cellular population growth and its exchange rates with the environment from metabolomics data, and estimation of promoter activity statistics from single-cell reporter gene data. In both cases, in the two separate papers addressing these problems (respectively, [24] and [22]), we propose solutions where stochastic process modelling enters the picture, and the problem is solved via a form of regularized estimation. However, the two problems pertain sharply different experimental scenarios and, as we will see, stochasting modelling enters in fundamentally different ways. These contributions constitute a good example of how tools from system and control theory may strongly contribute to the analysis and exploitation of biological data and, conversely, how modern microbiology experimental techniques can induce original developments in the control field. From a broader standpoint, they witness the important role of a data-processing interdisciplinary profile to enable the cross-talk between the control and the biology communities and help them to navigate through each other's domains.

3.2 Reconstruction of metabolic exchange rates from time-lapse metabolomics data

Modern metabolomics techniques allow for the monitoring of extracellular concentration of different metabolites over time-course microbial growth experiments [86]. Such time-resolved data enable one to address the reconstruction of time-varying (population growth and) metabolite exchange rates between the cellular population and the environment. In turn, knowledge of these exchange rate profiles constitutes the starting point for an in-depth analysis of intracellular metabolism and its rearrangements in response to environmental changes, such as enrichment or depletion of different carbon sources.

The apparent simplicity of reconstructing exchange rates from concentration profiles is deceptive. Measurements are noisy and sparsely sampled, making reconstruction ill-posed. Rate dynamics are qualitatively different over time, alternating slow fluctuations in constant metabolic regimes with sharp variations reflecting environmental changes and related metabolic reorganization (see Fig. 3.1 for an example). In addition, growth and exchange rates are mutually correlated. Appropriate account of these properties is needed to ensure that reconstructed rates are exploitable in further analysis.

In [24], we proposed a Bayesian regularization approach to the rate reconstruction problem that addresses all of the above concerns. For the simplest scenario where microbial growth is in batch, we rely on an Ordinary Differential Equation (ODE) model for biomass $b(t)$ (t denotes times) and extracellular concentrations $c_i(t)$ of n metabolites,

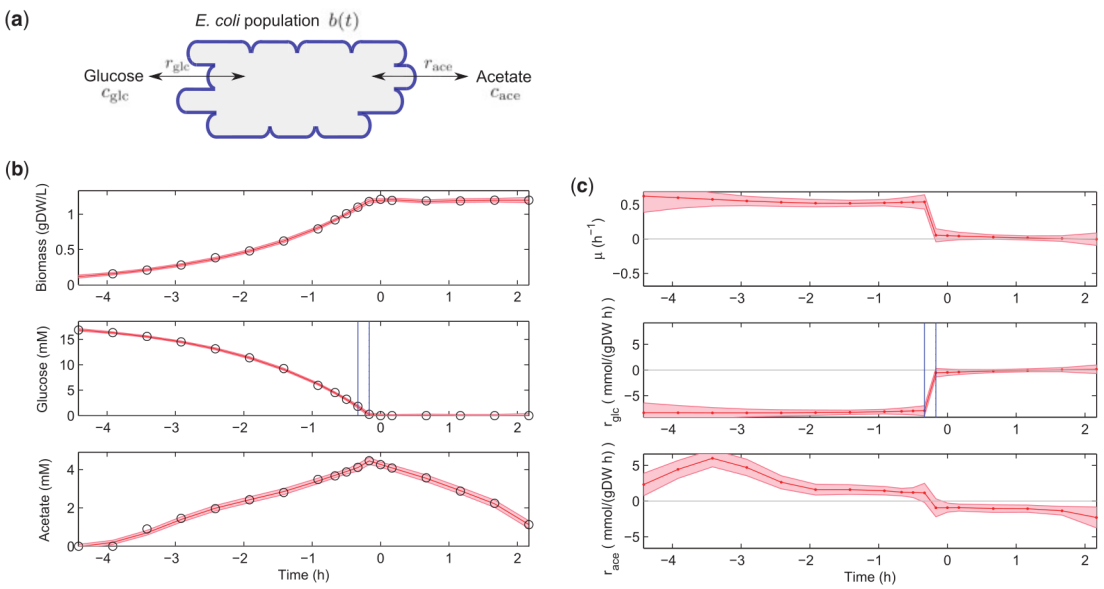

Figure 3.1: Reconstruction of growth and metabolite exchange rates from *E.coli* diauxic shift experiments with the method in [24]. (a) Illustration of the cell population and its metabolite exchange rates with the environment. (b) biomass and environmental metabolite concentration measurements (circles) and their estimation at all times. (c) growth and metabolite exchange rate estimates (vertical blue lines show the periods of fast transitions detected in data-preprocessing). In (b) and (c), estimates are in solid red, 95% credibility bands are in shaded red.

with $i = 1, ..., n$, of the form [31]

$$\begin{aligned} \dot{b}(t) &= \mu(t)b(t), \\ \dot{c}_i(t) &= r_i(t)b(t), \quad i = 1, \dots, n. \end{aligned} \tag{3.1}$$

Growth rate $\mu(t)$ and exchange rates $r_i(t)$ must be reconstructed over a whole experimental period $t \in \mathscr{T}$ from noisy measurements of b and of the c_i taken at possibly different time instants. The problem as it stands is clearly ill-posed [8]. In order to obtain a well-posed problem we use Gaussian process priors [98]. Namely we assume that every unknown rate r_i is the integral of a Wiener process, that is,

$$\ddot{r}_i(t) = \gamma(t)u(t), \tag{3.2}$$

with u a standard white noise process and γ a regularization factor that we will discuss further below (same for rate $\mu(t)$; u and γ are generally different across rates $\mu, r_1, ..., r_n$). By this modelling, the problem becomes that of computing the best estimates in a Bayesian sense, that is, provided a measurement model for the data $\mathscr{Y}$ (see measurement model in [24]), to calculate the a posteriori expectation

$$\mathbb{E}[\big(r_1(t), \dots, r_n(t), \mu(t)\big)|\mathscr{Y}] \tag{3.3}$$

at all times of interest $t \in \mathscr{T}$.

It can be shown that this problem formulation directly relates with Tikhonov regularization, with (3.2) acting as a curvature penalization term (see [118, 33]), thus favoring smooth solutions. Most importantly, we let regularization factor $\gamma(t)$ depend on time, such that larger values of $\gamma(t)$ can be used to capture steep transitions of the corresponding rate in proximity of sudden environmental/metabolism changes. As proposed in [24], $\gamma(\cdot)$ can be tuned in a data preprocessing step that isolates the time intervals where the transitions are likely to occur (measured substrate concentrations approaching zero, etc.). In practice, estimates (3.3) can be efficiently computed by an Extended Kalman Smoother (EKS, a forward recursion in the form of a standard Extended Kalman Filter, EKF, followed by a smoothing backward recursion [56]) for the augmented system made of the composition of (3.1) and the second-order linear (stochastic) dynamics (3.2) of every rate function. By construction, this provides credibility intervals (estimation error variance) along with the estimates.

Several comments are in order. A simpler, traditional approach would suggest to compute estimates of every rate by differentiating a spline fit of the corresponding concentration measurements. In such context, our time-dependent choice of $\gamma(\cdot)$ would roughly translate into appropriate collocation of the spline knots. This empirical approach has several drawbacks. Among other things, it does not account for the coupling of the different model equations, it has uncertain theoretical interpretation, quality of estimates needs to be assessed in further processing, and it is practically hard to handle. Though more complex, our approach provides automatic tuning of the smoothing profile and a simultaneous estimate of all rates at all times from noisy and sparse data, thus addressing all challenges raised by the problem. It is worth noting that the dynamical

formulation of regularized regression enabled by (3.2) is crucial for the effective calculation of (3.3). The main limitation of our implementation, possible divergence of the EKF, only affects the forward recursion. This can be ameliorated by replacing the EKF with more advanced nonlinear variants of the forward filter, such as Unscented Kalman or Particle Filtering [35, 55].

In [24], we demonstrated via simulation that this approach largely outperforms traditional approaches, such as the spline-based method mentioned above, on realistic synthetic datasets. As a side result, we also illustrated for this offline data processing problem the great improvement ensured by the smoothing step relative to the preliminary Kalman filtering results. We then applied the method on diauxic shift data from *E.coli* batch experiments (see Fig. 3.1), as well as on lactic acid production data from *L.lactis* fed-batch experiments. In the first case, validation of the estimates by metabolic flux analysis reconfirmed the relevance of the results even during sharp metabolic transitions. In the second case, inspection of the results quantified subtle metabolic regulation effects that are at present only qualitatively known, thus reconfirming the interest of our work. The tools used for the research were packaged into a basic Matlab software freely available for download and applicable to analogous metabolomics data analysis.

Developed in collaboration with an experimental biology group at INSA Toulouse [65], the method proposed in [24] is among the few existing ones for the estimation of time-varying rates from metabolomics time-profiles, and the first capable to provide accurate estimates at metabolic transitions in a completely automated manner. From the methodological viewpoint, the (auto-tuning) time-inhomogeneous approach to regularized inversion of dynamical systems is a strategy of general applicability that constitutes the key contribution of this work. My personal contributions consisted in both the main methodological solutions and the shared coordination of the project with Delphine Ropers (IBIS).

3.3 Generalized moment equations and inference of promoter activity statistics

In this section we look at an estimation problem stemming from the single-cell monitoring of gene expression. Following up from Section 1.2, one way to dynamically monitor gene expression at the level of single cells is to use fluorescent reporters co-expressed with the gene of interest. The fluorescence profile of a given cell is only an indirect account of the activation of the gene, since observations result from the promoter activity after gene transcription and translation are accomplished (see Fig. 1.2). In the study of gene regulation, however, the process of interest is precisely the promoter activation dynamics. A chief problem is thus to reconstruct (properties of) the latter from single-cell fluorescence measurements.

A common technique to quantify single-cell fluorescence dynamics in a population is by taking snapshots of fluorescence distributions in mutually independent cell samples

taken at different times. This can be the direct result of *e.g.* flow-cytometry measurements, or the outcome of videomicroscopy in absence of cell tracking [110, 47, 79]. In this scenario, rather than reconstruction from single-cell profiles, the problem becomes that of reconstructing promoter activity from time profiles of the fluorescence statistics in the population. Here, two challenges related with gene expression noise intervene:

1. The gene expression outcome of the same promoter activity in different cells is generally different;

2. The promoter activity in an identical environment may be different across cells.

Whereas the first challenge is essentially a modelling one, the second challenge alters the nature of the problem, since a unique promoter activity profile can no longer be even defined. An appropriate way to state the problem is instead to treat promoter activity as a stochastic process with different random outcomes in different cells, and to address reconstruction of the process statistics from the fluorescent statistics time profiles. Modelling promoter activity as a stochastic process itself is mechanistically well justified, since promoter regulation results from transcription factors that are the fruit of noisy expression of other genes. Thus, in sharp contrast with the previous section, stochastic modelling of the unknown is not chosen for mathematical convenience but results from the physics of the process. However, the details of this process are a priori unknown.

In [22], we consider the problem in the broader context of chemical reaction networks with rates (reaction probabilities in infinitesimal time) that are possibly affected by an exogenous (vector) process. We focus on reaction networks with reaction rate vector w that is affine in the network state X, that is, at time t,

$$w(t) = WX(t) + F(t), \tag{3.4}$$

where W is a known matrix (generalization to time-dependent W is also possible). For $F(t)$ a deterministic function of time, well-known moment equations provide in this case the time evolution of mean vector and covariance matrix of X in the form of a finite-dimensional system of ODEs ([48, 128]; the same applies for higher-order moments, which we do not consider). We instead treat $F(t)$ as a stochastic process. Under suitable hypotheses of absence of feedback (a nontrivial concept in the stochastic context, see discussion in the paper) we refer to F as an input process. For this case, we develop in [22] Generalized Moment Equations (GME) to relate the statistics of F with those of X. Under minimal assumptions on F (bounded mean and variance), we derive a system of ODEs that expresses second-order moments of X as a dynamical transformation of the second-order statistics of F. This ODE system includes standard moment equations as a special case. Most importantly, we show the transformation to be linear.

We then instantiate this framework for the gene expression problem of interest. Stochastic reporter expression dynamics are described by the so-called random telegraph model [87]. For a two-dimensional state $X(t)$ counting the number of mRNA and

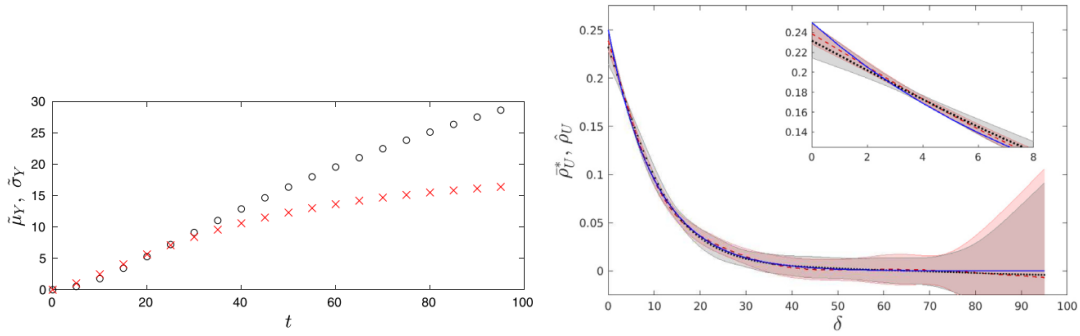

Figure 3.2: Reconstruction of promoter activity autocovariance function from fluorescent reporter mean and variance time-profiles (*in silico* results from [22]). (Left) noisy measurements of reporter mean (black circles) and standard deviation (red crosses) in a simulated population of cells. (Right) true promoter autocovariance function (blue) and its estimation statistics from 100 repeated experiments (dotted black line: estimation mean profile with automated choice of regularization factor ; dashed red line: same with manual choice of the same factor ; red and gray shaded regions: corresponding 95% confidence intervals).

protein molecules in a cell at time t, and $U(t)$ denoting the promoter activation state, this is a widely utilized model that describes the evolution of X in terms of stochastic transcription, translation and degradation reactions with transcription rate proportional to $U(t)$. This accounts for Challenge 1 above in terms of intrinsic noise (see next section for perspective generalizations). In turn, we propose to describe U as a stochastic input process via a corresponding definition of F (the hypothesis of absence of feedback is reasonable since fluorescence reporters do not act as regulators in the host cells). This choice effectively qualifies U as extrinsic noise and defines our approach to Challenge 2.

When applied to the random telegraph model of gene expression, the general results outlined above provide appropriate tools to address a variety of analysis and estimation questions. In [22], we focused on the problem of reconstructing promoter activity in the following sense (see also Fig. 3.2): We address inference of the statistics of U from the observed mean and variance profile of Y, with Y denoting the stochastic fluorescent reporter abundance (one of the entries of X). When restricted to the reconstruction of the mean profile of U from the mean fluorescence profile, it can be shown that the problem reduces to standard deconvolution, as addressed in theoretical as well as in a few methodological papers focused onn microbiology data [40, 105, 130]. When lifted to the space of second-order moments, we first of all notice that the variance of Y is crucially affected by the entire autocovariance function U. This is at the same time important and intriguing. Assuming (weak) stationarity of U, which can be justified for a variety of experiments, the autocovariance ρ of U describes in particular the memory of the promoter activity process. Thus, reconstructing ρ from data means learning about fundamental properties of promoter regulation without postulating anything about the mechanisms behind it (indeed U need not even be Markovian).

Because of the linear relationship between input and output statistics, reconstruc-

tion of ρ amounts to solving a linear inversion problem. Once again, mean and variance measurements of Y are typically noisy and sparse, which calls for regularized estimation. In [22], we addressed the problem by Tikhonov regularization, subject to the additional (convex) constraint that the solution must be found in the space of well-defined (positive semi-definite) autocovariance functions. For computational practicality, this is solved approximately in terms of a finitely-parametrized linear-quadratic program [14]. Via simulation, it was possible to evaluate the statistical properties of this estimator (Fig. 3.2), which is capable of satisfactory reconstruction of ρ in a realistic experimental setup. The problem addressed and the approach developed are quite novel. Indeed, in the literature, promoter statistics across different cells are either treated in the context of much more demanding experiments involving single-cell tracking, or are reduced to a common activation profile across all cells. I am the sole author of the entire work in context of the ANR project MEMIP [75], where further collaborative developments including applications to real experiments are being explored most notably in collaboration with Jakob Ruess at InBio [52].

3.4 Discussion and perspectives

We have looked in this chapter at two estimation problems of time-varying biological quantities from sampled data, namely reconstruction of cellular population growth and exchange rates from metabolomics data, and estimation of promoter activity statistics from fuorescent reporter population snapshot data. Connected in abstract terms, the problems are radically different in terms of experimental measurements and systems. The solutions proposed have been the object of two different publications. They cover a broad spectrum of applied methodologied from system and control theory, advanced statistics and machine learning.

The first problem has been addressed with application to real data, showing the interest and practicality of our solution and allowing us to already draw some original biological conclusions from the data analysed. While demonstrated on metabolomics data, the estimation methods developed in this work are of much more general applicability, for instance they could be applied to infer promoter activity from fluorescence data in the deterministic setting of Chapter 1. It is important to recall that the relevant application is accompanied by software that is not simply reproducing the publication results, but it can be applied to other experiments as described in an accompanying user guide. Publication of software implementing the published methods is a general and much desirable trend that reconfirms the interest of the community in such data analysis tools. Indeed, a few groups contacted us in connection with the usage of the software on own data. While the methodology proposed in the paper is rather mature, new versions of the software, focused in particular on robustification and inclusion of a graphical user interface for use by non-experts, are expected to be the object of an Inria technological development project proposal to be submitted in the near future with Delphine Ropers at IBIS.

The second problem has been treated in a more general mathematical context, and the resulting specific estimation algorithm has been tested on simulated data. By this we show the feasibility of reconstruction and the practicality of our solution. Yet, application to real data will certainly pose new challenges. Some of these concern the dependence of reconstruction performance on the system itself. Others concern possibly unmodelled dynamics, such as in the first place, sources of extrinsic noise that have not been accounted for in the reporter gene expression model pertaining Challenge 1. As anticipated above, this and other generalizations will be pursued in the context of project MEMIP [75].

Much broader applications of the framework developed in [22] are possible. In presence of data from single-cell tracking, exploitation of the empirical autocovariance function of the data is possible by appropriate modification of the estimation algorithm. This resonates the Mixed-Effects approaches discussed in Chapter 2, where the whole set of single-cell measurements is used at once for the direct estimation of a statistical population property. Moreover, further results reported in [22] concern the spectral analysis of reporter gene systems and show that, at a single-cell level, they can be seen as filters operating on the promoter activity profile and adding a noise component due to intrinsic gene expression noise. By this characterization, optimal design of reporter systems can be addressed in terms of spectral filter design that maximally preserves the input spectrum while minimizing the amount of added noise. Among other experimental design problems related with MEMIP, this design problem is meant to be the object of further analytical and experimental study with the experimental personnel of IBIS and BIOP [13] or other interested experimental groups.

It should be noticed that our approach allows one to decouple the reconstruction of statistical promoter activity properties from the study of the regulatory mechanisms that determine them. In addition, statistics of different promoters can be inferred in different experiments devoted to the monitoring of the corresponding genes. Unlike single-cell traces, population statistics can be directly compared across experiments. One may thus think to further exploit the statistics obtained for different promoters toward a stochastic investigation of the regulatory interactions among the observed genes. This corresponds to lifting the gene regulatory network reconstruction problem from the deterministic setup treated in Chapter 1 to a potentially much more informative stochastic setup, with promising impact on the understanding of regulatory networks. Initiated in [21], this line of research shall be taken further in new projects to be defined.

[24] Estimation of time-varying growth, uptake and excretion rates from dynamic metabolomics data (2017). E.Cinquemani, V.Laroute, M.Cocaign-Bousquet, H.de Jong, D.Ropers, *Bioinformatics* (Proceedings of the 25th ISMB/16th ECCB), 33(14):i301–i310.

Abstract: *Motivation.* Technological advances in metabolomics have made it possible to monitor the concentration of extracellular metabolites over time. From these data, it is possible to compute the rates of uptake and excretion of the metabolites by a growing cell population, providing precious information on the functioning of intracellular metabolism. The computation of the rate of these exchange reactions, however, is difficult to achieve in practice for a number of reasons, notably noisy measurements, correlations between the concentration profiles of the different extracellular metabolites, and discontinuties in the profiles due to sudden changes in metabolic regime. *Results.* We present a method for precisely estimating time-varying uptake and excretion rates from time-series measurements of extracellular metabolite concentrations, specifically addressing all of the above issues. The estimation problem is formulated in a regularized Bayesian framework and solved by a combination of extended Kalman filtering and smoothing. The method is shown to improve upon methods based on spline smoothing of the data. Moreover, when applied to two actual datasets, the method recovers known features of overflow metabolism in *Escherichia coli* and *Lactococcus lactis*, and provides evidence for acetate uptake by *L. lactis* after glucose exhaustion. The results raise interesting perspectives for further work on rate estimation from measurements of intracellular metabolites.

[22] Stochastic reaction networks with input processes: Analysis and application to gene expression inference (2019). E.Cinquemani, *Automatica*, 101:150-156.

Abstract: Stochastic reaction network modelling is widely utilized to describe the probabilistic dynamics of biochemical systems in general, and gene interaction networks in particular. The statistical analysis of the response of these systems to perturbation inputs is typically dependent on specific perturbation models. Motivated by reporter gene systems, widely utilized in biology to monitor gene activity in individual cells, we address the analysis of reaction networks with state-affine rates in presence of an input process. We develop a generalization of the so-called moment equations that precisely accounts for the first- and second-order moments of arbitrary inputs without the need for a model of the input process, as well as spectral relationships between the network input and state. We then apply these results to develop a method for the reconstruction of the autocovariance function of gene activity from reporter gene population-snapshot data, a crucial step toward the investigation of gene regulation, and demonstrate its performance on a simulated case study.

Chapter 4

Control of microbial growth and microbial communities

4.1 Natural and synthetic microorganism control

The concept of control enters the study of microbial dynamics in different flavors. Often mentioned in the previous sections, life of a cell is sustained by a complex set of biochemical regulatory interactions enforcing crucial dynamics such as the cell cycle, adaptation to changing environments, and so on. In natural environments, evolution has driven the selection of organisms based on their ability to grow and outcompete other species in a shared environment. In a first assessment, growth of microorganisms can thus be seen as an optimal control problem that shaped internal resource allocation regulatory mechanisms toward maximal growth rate or total biomass production [84, 31, 44].

Recent advances in genome engineering have enabled the directed modification of the metabolic pathways and regulatory networks of the cell at an unprecedented scale, thus enabling microorganisms to be turned into microbial cell factories capable of producing a range of metabolites, peptides, and proteins of industrial and medical interest [54]. For a suitably engineered species, a consequent control challenge is to determine the best control inputs (*e.g.* in the form of biochemical perturbations or light stimuli) and feedback strategies from real-time measurements that optimize the synthesis of the target product [125].

In nature, microbial populations rarely occur in isolation, but rather form communities in a shared environment [93]. They are thus subject to mutual interactions giving rise to a richer set of behaviors than the single species [38]. Engineering efforts have also been extended to this level by the development of synthetic microbial consortia [7]. Today, exploitation of natural or synthetic microbial communities is being pursued in a vast range of scenarios, for instance, applications in the biotechnology and pharmaceutical industries [122]. Still, most often, microbial communities are regarded as an unresolved whole, without a detailed account of their inherent heterogeneity.

Combining approaches from the modelling of microbial growth dynamics and ecology, new mathematical tools need to be developed for the analysis of the complex interaction dynamics characterizing microbial consortia [59, 49, 46, 94, 39]. From the experimental viewpoint, compared with natural communities, synthetic microbial communities offer a more favorable ground for the study of the interaction dynamics, since they are better circumscribed, easier to manipulate and monitor. Under the premises of appropriate modelling and experimental tools, optimal control of microbial consortia is expected to outperform single-species control in applications of practical relevance [104, 7].

The control problems discussed above refer to deterministic approaches for population-average dynamics. In presence of significant cell response variability, or for single-cell control problems [18], this approach may be insufficient. A control approach based on stochastic cell-response models (see Chapter 2 and Section 3.3) may be considered instead. Based on these models, by appropriate theoretical tools from stochastic optimization, control of population mean response may be addressed in conjunction with minimization of the dispersion (variance) of the response. As another example, single-cell control problems addressed in the literature with deterministic control tools [114, 76] could be better addressed in a stochastic setup, requiring for instance that cell response remains in a given range with preassigned probability.

In this chapter, we discuss control problems both in terms of resource allocation in a single species [25] and in terms of interactions in a synthetic consortium [73]. An important point that applies to both problems is the focus on minimal phenomenological models. This is not only motivated by ease of mathematical treatment and calibration with data, but also, in a control perspective, by the need to design model-based controllers and implement them in real time. The results are in both cases from the early stages of my research activity in this direction, which explains the preliminary nature of the papers reviewed. Finally, we make a detour into the formulation of control problems for stochastic systems, discussing one contribution in stochastic optimal control in presence of probabilistic constraints [23]. The type of tools illustrated by this theoretical contribution could later be applied to concrete single-cell, cell-population, or microbial consortium control problems.

4.2 Resource allocation control for optimal bacterial growth and productivity

Laboratory studies and modelling of microorganisms often explore one or several microbial growth regimes without a detailed investigation of transitions among them. However, in nature, the time microorganisms spend in steady-state (in the biological sense of constant growth rate) regime may even be negligible compared with quiescence or transition periods such as growth arrest, restart, and metabolic reorganization [44]. In response to this observation, thanks to dynamic gene expression monitoring and other time-lapse experimental techniques, increasing research efforts are devoted to the study of resource (re)allocation in bacteria in response to environmental changes such as sud-

den availability/depletion of a sugar (also see the discussion of Section 3.2).

An especially promising approach is the usage of coarse-grained models describing resource allocation dynamics in terms of few aggregated quantities, such as rybosomes for genetic machinery, enzymes for the metabolic machinery, etc., and fundamental regulatory effects among them (see Fig. 4.1). In this approach, resource allocation is captured by a factor $\alpha(t)$ that regulates over time t the balance of uptaken substrates directed to one or the other compartment in response to environmental changes. Intentionally simple, this modelling of cellular dynamics lends itself to analytical or semi-analytical investigation of the allocation strategy $\alpha(\cdot)$.

In view of evolutionary arguments, a well-accepted criterion is that microorganisms are selected to maximize their ability to grow. This criterion is used *e.g.* in Flux-Balance analysis, where reaction rates of underdetermined data fitting problems are often sought among the solutions maximizing a growth rate expression [84]. In the context discussed here, this suggests to investigate optimal resource allocation strategy $\alpha(\cdot)$ as the solution of a suitable variational optimization problem. By a similar rationale, in a synthetic biology spirit, one may instead investigate how this resource allocation strategy should be altered to optimize a different artificial objective, for instance, maximal synthesis of a product of interest.

In previous studies [44, 125], our group addressed these questions in terms of two different optimization problems for maximal biomass production,

$$\max_{\alpha(\cdot)\in\mathscr{U}} \log \frac{\mathscr{V}}{\mathscr{V}_0} = \int_0^T \mu(t)dt, \tag{4.1}$$

and maximal biochemical product synthesis,

$$\max_{\alpha(\cdot)\in\mathscr{U}} X(T) = \int_0^T V_X(t)dt, \tag{4.2}$$

where, importantly, $\mathscr{U}$ is the space of functions bounded between 0 and 1. Growth rate μ, total biomass $\mathscr{V}$ (equal to $\mathscr{V}_0$ at $t = 0$), and synthesis rate V_X of the target product with total abundance X all depend on α via a simple ODE system (see [44, 125, 25]). For the first problem, analytical study of the problem via the Pontryagin maximum principle and differential inclusion tools allowed authors to conclude that the solution is a bang-bang-singular strategy (similar to what illustrated in Fig. 4.1), with bang-bang regimes associated with fast metabolic reorganization in correspondence of sudden substrate concentration changes, and a singular regime in-between corresponding to steady-state growth and biomass synthesis. Qualitatively speaking, this oscillatory strategy was biologically explained in terms of known regulatory modules. Similar mathematical conclusions were drawn also by the aid of simulation for the second problem.

From the quantitative biological viewpoint, a bang-bang solution for resource allocation (with infinitely many transitions around an accumulation point) is not realistic. Resource reallocation certainly obeys nontrivial dynamics and, most importantly, must come at a price (energy expenditure, etc.). From the mathematical viewpoint, given the

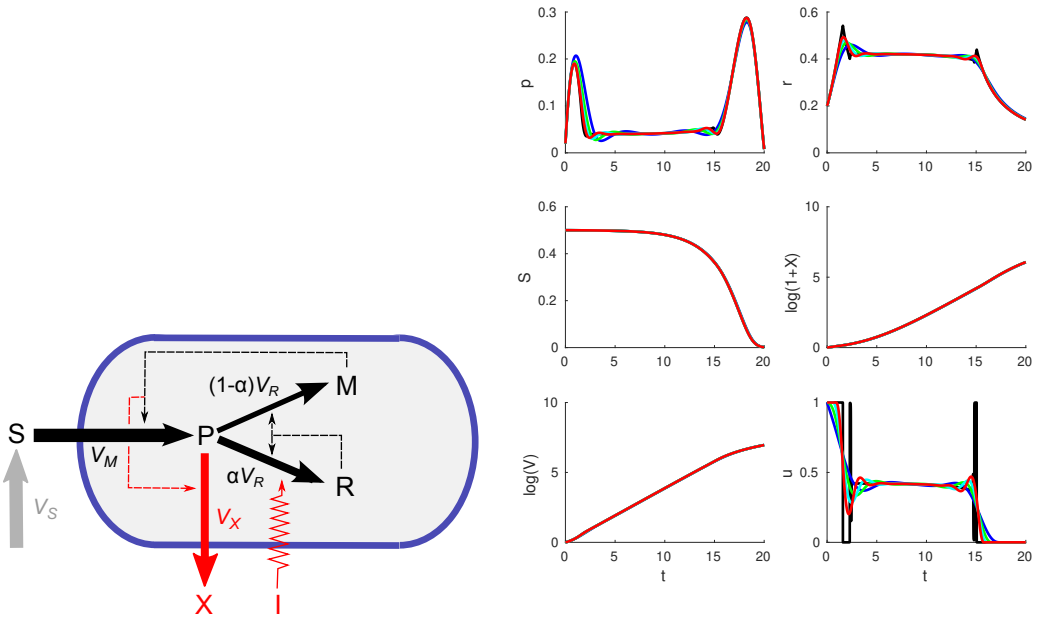

Figure 4.1: Illustration and analysis of the resource allocation model of [125, 25], including synthesis of a target molecule X. (Left) In a prototypical 'average' cell (gray rod-shaped area with blue bounds), M denotes metabolism-related enzymes, R denotes gene expression machinery (rybosomes), V_M is the uptake rate of substrate S (possibly added to the medium at a rate V_S) yielding precursors P, V_R is the precursor utilization rate, α is the balance of utilization of P for M or R synthesis. Additionally, in red, controlled synthesis of X at a rate V_X under external control I. Thin arrows denote regulatory effects, thick arrows denote fluxes. (Right) System dynamics from the solution of the maximal product synthesis optimization problem (4.2) (black) and of its regularized variants with different weights λ of the regularization factor (4.3) (other colors; in the bottom-right panel, u accounts for α in presence of the exogenous control input I).

structure of $\mathscr{U}$, a switching solution cannot be avoided unless smoothness is explicitly enforced. A more relevant, and perhaps even more convenient, formulation of the two problems is proposed in [25]. This amounts to the addition in the objective function of a term that penalizes fast variations of α, of the form

$$-\lambda \cdot \int_0^T w(t)^2 dt, \tag{4.3}$$

with $\lambda > 0$, where $w = \ddot{\alpha}$. Borrowed from regularized estimation theory, the approach ensures smoother solutions for larger values of λ, at the price of suboptimality of the original problem (In practice, for a correct formulation and numerical practicality, the optimization is expressed in terms of w).

In [25], we show by numerical investigation that this problem formulation returns smooth oscillatory solutions that approach bang-bang solutions with $\lambda \to 0$ (see Fig. 4.1). For strictly positive λ, biologically realistic (smooth) solutions are found that do not significantly alter the predicted intracellular dynamics and the attained values of the original objective function. In addition, a clear understanding of the role of time horizon T could be determined (see details in [25]).

In sums, the new approach allows us to explore the resource allocation problem on a more plausible biological ground in a simpler analytical framework. Of course, more work is needed to pinpoint the biological origin and the most appropriate form of cost (4.3). Following up from previous work of the group and collaborators, my personal contribution was the formulation of a regularized optimization problem and of a convenient numerical implementation, and the analysis of the results.

4.3 Toward feedback control of synthetic microbial communities

By their very essence, microbial communities make up a system with more complex dynamics than one species alone [38, 108, 46]. Different species exert mutual regulation by the exchange of and competition for substrates in the environment, let apart direct regulation such as biochemical messages, quorum sensing, and the like [127]. As a result, in nature, microbial communities perform tasks that any of the species of the community could not perform in isolation, to the profit of all species [38, 59, 58].

A common and simple interaction pattern is one where a first species, growing on a substrate and excreting a toxic byproduct, is accompanied by a second species that instead exploits this byproduct for its own growth, with the beneficial effect of cleaning the environment for the first. This type of ecosystem occurs in nature, for instance, among different variants of *E.coli* bacteria [111, 58]. Analogous synergies also occur among different microorganisms, *e.g.* in the digestive system of mammals [80, 93, 94].

Also learning from natural evolution, it is intuitive to think that microbial communities can be exploited to perform tasks of societal interest better than single species

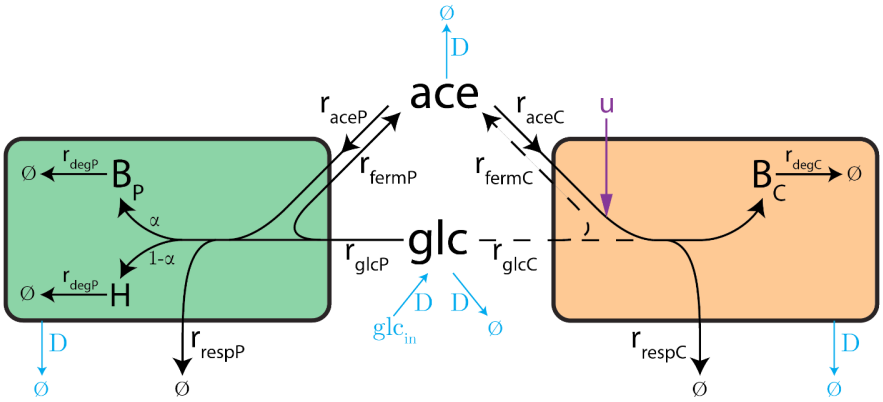

Figure 4.2: Illustration of the *E.coli* producer-cleaner microbial community model in a bioreactor environment to be published in [73]. Dynamics of the producer population (green, with total self-replicative biomass B_P) and of the cleaner population (orange, with total biomass B_C) are determined by several glucose (glc) and acetate (ace) exchange rates with the environment and loss rates (black lines), that generally differ between the two in consequence of suitable genetic engineering. Glucose uptake of the producer is redirected from growth to synthesis of a heterologous protein H in a (fixed) percentage α. Cyan denotes quantities and fluxes pertaining bioreactor dynamics (dilution rate D, input flux glucose concentration $\mathrm{glc}_{\mathrm{in}}$), purple is for the effects of exogenous control u.

alone. In the context of a broader project (we will come back to this project in the following discussion section and in Chapter 5), we have started to investigate the synthesis of a protein of interest by means of a synthetic microbial consortium. The objective is to study coexistence dynamics and implement optimal control strategies achieving increased productivity relative to a single producer species.

We are working on a system based on *E.coli* bacteria precisely implementing the interaction pattern outlined above (see Fig. 4.2). A first wild-type *E.coli* population grows on glucose and produces acetate as a byproduct. In high concentrations, acetate is toxic for growth on glucose. A second, modified variant of *E.coli* bacteria forms a second population that instead grows primarily, but less efficiently, on acetate, also consuming glucose but to a lesser extent. In addition, the first species may be itself engineered so as to allocate some resources away from growth to the synthesis of a target (heterologous) protein H. We refer to the first species as producer and the second as cleaner.

For the sake of H synthesis, a number of tradeoffs are apparent. How much resources should the producer steer away from growth to production, such that the cumulated productivity of the whole biomass is maximal ? How many cleaners should be present such that the production gain from detoxification is larger than detrimental competition for glucose ? Answering these questions quantitatively and in the most general case of a dynamical environment requires mathematical models of microbial communities that are only starting to be developed [59, 49, 46]. In a control perspective, provided the engineering of genetic circuits responsive to external induction and suitable monitoring

of the cellular populations, additional challenges come from the design of viable, model-based optimal control laws.

With reference to Fig. 4.2, we have developed an ODE model for the time evolution of the growing biomass of the producer, B_P, and of the cleaner, B_C, of the synthesized protein H, and of the environmental concentrations of glucose (glc) and acetate (ace) in a fixed reaction volume,

$$\dot{x} = f\big(x, D, glc_{in}, u\big), \qquad x = (B_P, B_C, H, ace, glc), \qquad x(0) = x_0. \tag{4.4}$$

Here, u is an exogenous control input that we assume to regulate the cleaner uptake rate of ace and thus the resulting cleaner growth rate (all quantities potentially depend on time). The model also accounts for possible input and output flows at equal (dilution) rate D, with glucose added in the input flow at concentration glc_{in}. In particular, equations for B_* ($*$ indicating any of the two species) are of the form $\dot{B}_* = \mu_*(x) \cdot B_* - D \cdot B_*$, where μ_* is by definition the growth rate.

As emphasized in the figure, the model is expressed in terms of the exchange rates between cellular populations and environment. These are state-dependent phenomenological laws capturing the known essential metabolic dynamics (metabolism overflow determining ace excretion, toxicity of high ace concentrations, etc.). Compared with the resource allocation model of Fig. 4.1, internal cellular dynamics are here entirely abstracted away, while the complexity of the model resides in the resulting interaction dynamics (and the dependence on u). The model has been calibrated based on rate measurements from the literature, and was found capable to quantitatively reproduce relevant steady-state as well as time-course data.

Based on this model, we want to investigate whether the presence of a cleaner may enhance production of H. One way to formulate this question is in batch, that is, assuming $D = 0$ and a fixed amount of glucose initially provided to the culture. This leads to an optimization problem of the type

$$\max_u H(T) \text{ subject to (4.4)}, \tag{4.5}$$

where $H(T)$ is the total amount of protein H at a final time T. The solution to this problem can then be compared with the value of $H(T)$ obtained from a reduced model for the sole producer. However, this is not a firm ground for comparison, due to the dependence on initial conditions and the time horizon T.

A better choice is to first perform the analysis in chemostat, that is, for a constant (time-invariant) u, a constant $D > 0$, and a constant $glc_{in} > 0$, pursuing steady-state analysis of (4.4). The standard form (primarily Michaelis-Menten and threshold functions) of reaction rates is such that vector field f is composed of piecewise rational functions over a partitioning of the state space. Therefore, the search for equilibria by the equation $0 = f(x)$ can be reduced to the search of the (real, nonnegative) roots of a piecewise multinomial system of equations. Though the problem remains difficult, the search of multinomial roots can profit from extensive literature and specialized numerical tools. Results can be checked in simulation as well.

Several conclusions could be drawn from the analysis outlined above. A preliminary observation is that, in steady-state, existence of a cellular population, expressed by $B_* > 0$, is possible only if $\mu_* = D$. Coexistence in steady state of both producers and cleaners thus implies the ability to grow at equal rate D, in spite of the fact that the maximal producer growth rate is larger than the cleaner counterpart. With a mild dependence on glc_{in}, coexistence could be predicted for intermediary values of D, guaranteeing sufficient acetate excretion by the producer to sustain cleaner growth, but not exceeding the maximal cleaner growth rate. Second, larger proportions $1-\alpha$ of resources steered to H synthesis generally enlarge the range of values of D where coexistence is possible. Intuitively, this is because producer growth rates get reduced to values closer to cleaner growth rates. Coming to the original question of productivity, we found that the maximal steady-state extraction rate $\max_D H \cdot D$ is larger when cleaners are present than in a bioreactor with producers only (as described by a reduced version of (4.4)). This, however, comes at the price of a worse ratio $(H \cdot D)/(glc_{in} \cdot D)$, which represents production efficiency (product extraction rate per glucose injection rate), since part of the glucose fed is consumed by the cleaner.

Additional results were derived in alternative scenarios, reconfirming the productivity advantage that one may obtain from the microbial community in place of the sole producer species. These results will constitute the work presented in a paper under preparation [73]. My personal contributions to this work are in theoretical microbial community design, modelling and model analysis, in addition to the co-supervision of the work of Marco Mauri (first author of the work) in his PostDoc stay at IBIS. Results developed so far provide the necessary mathematical basis for the investigation of model-based controller design, as discussed in Section 4.5.

4.4 Optimal constrained control of stochastic systems

Different from control of deterministic systems, stochastic optimal control deals explicitly with systems subject to disturbances or with inherent variability in their dynamics. Stochastic control is thus naturally suited to a variety of uncertain systems, included *e.g.* control of microorganisms at the single-cell level. From the viewpoint of classical control theory, a nice discussion of the theoretical advantages provided by stochastic control over deterministic control can be found in [5]. However, the theoretical and practical solution of stochastic control problems is generally more challenging. To illustrate the intricacies of optimal control of stochastic systems, let us start from the statement of a deterministic control problem. Consider for instance a discrete-time dynamical system

$$x(t+1) = f\big(x(t), u(t)\big), \qquad t = 0, \dots, N-1, \qquad x(0) = x_0, \tag{4.6}$$

where $x(t) \in \mathbb{R}^n$ is the state vector at time t, with given initial condition x_0, $u(t) \in \mathbb{R}^m$ is an input, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a deterministic state update map. Let $\bar{x} = \big[x(0)^T \cdots x(n)^T\big]^T$ and $\bar{u} = \big[u(0)^T \cdots u(n-1)^T\big]^T$. In great generality, optimal

(constrained) control can be written as the minimization problem

$$\min_{\bar{u}} \; V(\bar{x}, \bar{u}) \text{ subject to (4.6) and} \tag{4.7}$$

$$\eta(\bar{x}, \bar{u}) \leq 0, \tag{4.8}$$

where $V : \mathbb{R}^{(N+1)n \times Nm} \to \mathbb{R}$ is some cost function that expresses control objectives, and function $\eta : \mathbb{R}^{(N+1)n \times Nm} \to \mathbb{R}^r$ represents r possible constraints (inequality to be intended componentwise). Among these problems is Linear-Quadratic Regulation (LQR), where f is a linear map and V is a quadratic function [2]. In absence of constraints (4.8), LQR has a solution that can be expressed as a state feedback, that is, $u(t) = K(t)x(t)$, for some control gain $K(t)$ easily obtained via dynamic programming [2]. Note that knowledge of $x(t)$ at all times was not assumed to solve (4.7). Indeed optimization is expressed in the space of input trajectories, that is, it is an open-loop control problem. Even if applying a feedback control strategy has well known advantages, on the ground of (4.7), the open-loop solution and its closed-loop expression are equivalent due to the determinism of the system. In presence of constraints, as long as η is a convex (in particular, affine) function, the problem remains convex. Analysis of the problem is thus possible, and numerical solution is easy. For a practical feedback solution, a well-studied approach to constrained LQR is (linear) Model-Predictive Control (MPC, [100]).

In the stochastic setup, f is not a deterministic but a stochastic mapping. Abusing notation, one considers systems of the form

$$x(t+1) = f\big(x(t), u(t), w(t)\big), \qquad t = 0, \dots, N-1, \qquad x(0) = x_0, \tag{4.9}$$

where $w(t)$ is a (discrete-time) stochastic process. For these systems, problem (4.7)–(4.8) does not make sense in practice, since different problems should be solved depending on the (a priori unknown) fate of $w(t)$. A common and appropriate reformulation is in terms of the expected cost $\mathbb{E}[V(\bar{x}, \bar{u})]$. However, minimization of this cost with respect to $\bar{u}$ is wasteful. It can be shown that, unlike the deterministic case, seeking the solution in a space of trajectories is different from seeking the solution in a space of policies, that is, functions of the available observations of the system [5]. Thus, assuming the system state is observable, a more reasonable problem statement is

$$\min_{\theta \in \Theta} \; \mathbb{E}[V(\bar{x}_\theta, \bar{u}_\theta)] \text{ subject to (4.9) and} \tag{4.10}$$

$$\mathbb{E}[\phi \circ \eta(\bar{x}_\theta, \bar{u}_\theta)] \leq 0, \tag{4.11}$$

where $\{\bar{u}_\theta : \; \theta \in \Theta\}$ is a parametric family of functions of $\bar{x}_\theta$ and (4.9) holds with reference to $\bar{u}_\theta$ and $\bar{x}_\theta$ in place of $\bar{u}$ and $\bar{x}$ ($\bar{x}_\theta$ depends on θ via $\bar{u}_\theta$). Introduced in [23], for different definitions of the mapping $\phi : \mathbb{R}^r \to \mathbb{R}^R$, the general expectation-type constraint formulation (4.11) encompasses a variety of possible probability-type constraints (see [23] and further below), essentially aimed at enforcing the original hard constraint (4.8) for most of the random outcomes of w. In absence of constraints, the equivalent of LQR is the Linear Quadratic Gaussian (LQG) problem [5]. In LQG, V is again a quadratic objective function, w is a Gaussian process and f is linear in x,

u, and w. Among the causal policies, the solution of the LQG problem is the same state-feedback policy found in LQR.

The situation becomes more involved in presence of constraints. Convexity of the problem (4.10)–(4.11) depends not only on the expressions of V and η, as is the case, at least for linear dynamics, for the deterministic problem (4.7)–(4.8). It also crucially depends on the structure of the policies $\bar{u}_\theta$ and of the function ϕ that determines in what probabilistic sense the constraints have to be interpreted. Lack of convexity is a serious issue. This is because of absence of guarantees on the existence and uniqueness of the solution, and, even if a unique solution exists, it is extremely hard to determine both by analytical and numerical means, due to possible existence of local minima. Whereas convexity of (4.10) is relatively easy to guarantee, convexity of relevant instances of (4.11) is generally difficult to ensure.

In [23], we have studied stochastic optimal control problems of the form (4.10)–(4.11). In the common (and well-motivated) framework of disturbance-affine control policies, we have studied the convexity of the constraints (4.11). Based on the current literature on stochastic (model-predictive) control (see references in [23]) and on theoretical results from stochastic programming [92], we have proposed tractable methods for the convex approximation of several instances of the problem (different forms of η and ϕ). Specifically, we proposed approximations for $\mathbb{P}[\eta(\bar{x}_\theta, \bar{u}_\theta) \leq 0] \geq 1 - \alpha$ (chance constraints), with η a polytopic or ellipsoidal function, and, given any convex function η, for instances of (4.11) of the type $\mathbb{E}[\phi \circ \eta_i(\bar{x}_\theta, \bar{u}_\theta)] \leq \beta_i$, $i = 1, \ldots, R$, with ϕ a ramp function (so-called integrated chance constraints). Numerical simulations demonstrating the effectiveness of our solutions compared with existing ones accompanied the analytical treatment.

The work was carried out as a side project of my PostDoc at ETH. My personal contribution to the results reported in [23] was in terms of both analytical developments and numerical simulations, as well as the finalization of the project for publication.

4.5 Discussion and perspectives

We have discussed in this chapter several problems related with control of biological systems. The intracellular resource allocation problem has been addressed to understand natural strategies as a first step toward engineering of cell factories. From the viewpoint of natural strategies, the ability to cope with changing environments has shaped regulatory responses of living systems. Linking back to Section 1.2, the analysis of resource reallocation in transient conditions can also be seen as first step toward a finer understanding of regulatory networks and how they guarantee adaptation to environmental changes.

Despite encouraging experimental observations, more work has to be done to establish the accuracy of our results. A difficult point of the experimental validation of the predicted reallocation strategies is the ability to monitor resource reallocation with the necessary accuracy. One possible approach is the usage of fluorescent re-

porters for the monitoring of genes tightly coupled with resource allocation. However, de-synchronization of cells across the population may easily lead to population-average dynamical responses where the oscillatory behavior of the response is entirely lost. In view of this, single-cell analysis and monitoring could be necessary, and tools such as those discussed in Section 3.3 become of primary importance.

From the viewpoint of optimal control, modelling and analysis efforts shall be coupled with experimental activity devoted to the design and engineering of synthetic circuits for the external control of cellular dynamics. A key effort in this sense is the deployment of optogenetics, that is, genetic circuits responsive to light, which is effectively the state of the art for cell control [18, 102]. These aspects are part of a currently running ANR project I participate in, MAXIMIC [74]. In the framework of this project, research is being conducted in collaboration with the Inria project-team BIOCORE [12] and with the BIOP experimental personnel [13].

Optimal control of synthetic microbial communities has been discussed so far in terms of dynamical modelling of microbial interactions and the assessment of key properties of the producer-cleaner community we are constructing. We have developed a model that is capable of reproducing a rich set of data from the literature, and that lends itself to analytical or semi-analytical investigation. We have predicted the possibility of stable coexistence in chemostat conditions and the productivity gain associated with the presence of a second species. On the basis of these results, control design will be pursued with several objectives. A first objective is the design of control strategies aimed at maintaining optimal production regimes in the sense of the steady-state analysis above, in the face of environmental perturbations and modelling inaccuracies. A second objectives will be the study of optimal control strategies by time-varying production profiles, *e.g.*, by the solution of problems of the type (4.5). Investigation of optimal control problems will profit from my own expertise as well as that of additional collaborators, notably the Inria groups BIOCORE and VALSE [115].

The other important aspect of this research is the construction and experimental characterization of the microbial community, as well as the experimental implementation and validation of the theoretical models and the control strategies developed. This is the subject of a Ph.D. project that is expected to start soon under my co-supervision and the direction of Johannes Geiselmann. This project is in the context of the Inria IPL COSY [29], a larger project entirely devoted to synthetic control of microbial communities, which will be further discussed in Chapter 5. Via related technological development projects [28, 83], this includes the development of an automated experimental platform for the computer-based optogenetics feedback control of microbial communities. Developed in collaboration with BIOCORE, the engineering staff of Inria Sophia Antipolis (where BIOCORE is located) and the BIOP personnel, the platform will be used for characterization and control experiments related with both problems discussed in this chapter.

Finally, together with relevant developments in the control theory literature, the general tools for stochastic optimal control problems presented in the chapter admit intriguing applications to single-cell control. Control of individual-cell response based on

stochastic modelling is indeed in the scope of the IPL project that I coordinate [29]. Also leveraging recent developments in simulation-based controller design (*e.g.* the so-called scenario approach [16]), this represents a promising research avenue to be investigated with the IPL consortium, notably with team COMMANDS [27] and InBio [52].

[25] Optimal control of bacterial growth for metabolite production: The role of timing and costs of control (2019). E.Cinquemani, F.Mairet, I.Yegorov, H.de Jong, J.-L. Gouzé, *Proceedings of the 17th European Control Conference* (ECC 2019), to appear.

Abstract: The growth of microorganisms is controlled by strategies for the dynamical allocation of available resources over different cellular functions. Synthetic biology approaches are considered nowadays to artificially modify these strategies and turn microbial populations into biotechnological factories for the production of metabolites of interest. In our recent work,we have studied dynamics of microbial resource allocation and growth in terms of coarse-grained self-replicator models described by ordinary differential equations, and proposed artificial control strategies for the optimization of metabolite production based on the reengineering of resource allocation. In this paper, we elaborate on our earlier results and further investigate synthetic resource allocation control strategies. Using numerical simulation, we study the effect on growth and bioproduction of the (biological or technological) costs associated with discontinuous control strategies, and of the time allotted to optimal substrate utilization. Results provide novel insight into the most favorable synthetic control strategies.

[73] Modelling and design of a synthetic microbial community for enhanced productivity in bioreactor (2019). M.Mauri, J.L.Gouzé, H.de Jong, E.Cinquemani, in preparation.

[23] Convexity and convex approximations of discrete-time stochastic control problems with constraints (2011). E.Cinquemani, M.Agarwal, D.Chatterjee, J.Lygeros, *Automatica*, 47(9):2082–87.

Abstract: We investigate constrained optimal control problems for linear stochastic dynamical systems evolving in discrete time. We consider minimization of an expected value cost subject to probabilistic constraints. We study the convexity of a finite-horizon optimization problem in the case where the control policies are affine functions of the disturbance input. We propose an expectation-based method for the convex approximation of probabilistic constraints with polytopic constraint function, and a Linear Matrix Inequality (LMI) method for the convex approximation of probabilistic constraints with ellipsoidal constraint function. Finally, we introduce a class of convex expectation-type constraints that provide tractable approximations of the so-called integrated chance constraints. Performance of these methods and of existing convex approximation methods for probabilistic constraints is compared on a numerical example.

Chapter 5

Conclusions and outlook

In this manuscript I have provided an overview of my past and current research at the intersection of systems biology, estimation and control theory. The material has been arranged into four chapters built around representative publications that I (co-)authored, corresponding to different problems and methodological approaches. As discussed in each of the four chapters, the work has been developed in the context of several research projects, with the collaboration of a number of colleagues in France and abroad, and the (co-)supervision of Ph.D. students and PostDoc researchers. Additional research activities that have not been discussed pertain long-standing collaborations with Zoi Lygerou at University of Patras (Greece) and Maria Anna Rapsomaniki at IBM Zurich (Switzerland) [96], as well as further collaboration with John Lygeros at ETH Zurich (Switzerland) and Giancarlo Ferrari-Trecate (University of Pavia, Italy, now EPFL Lausanne, Switzerland) [90].

Despite the proposed classification, the different topics discussed in the previous chapters are clearly not independent from one another. Several connections among them have also been outlined. In a more unifying view, two broad subjects constitute my current activity and lay the ground of my future research directions, as follows.

Estimation of biological signals and systems. The ever-developing experimental techniques for the quantitative, dynamical monitoring of cellular dynamics produce rich and heteregeneous datasets. Dedicated mathematical analysis is necessary to keep in pace with these developments and ensure optimal utilization of the data. In my current and upcoming research, a great emphasis is placed on system identification and estimation of hidden dynamics from experimental data. Determining to what extent the dynamics of a biological system can be pinpointed from data is a fundamental problem toward the understanding of the underlying regulatory mechanisms. Identifiability studies spur a variety of questions, such as optimal experiment design, ensemble modelling, and network reconstruction, that are all part of my interests.

These challenges are especially open at the level of single-cell dynamics. At present, I am currently addressing some of these challenges in project MEMIP [75], in which I am

principal investigator for Inria Grenoble – Rhône-Alpes. Funded by the French National Research Agency (ANR), the project also involves the experimental groups InBio [52] (Gregory Batt) and MSC [78] (Pascal Hersen), and the applied mathematics group XPOP [124] (Marc Lavielle). In addition to the principal investigators, on these subjects I am collaborating with Aline Marguet, hired at IBIS as PostDoc and now freshly appointed Inria Research Scientist, as well as Jakob Ruess at InBio. Current aims are to further develop gene expression modelling over branching cellular popolations (see Section 2.3) to include intrinsic noise, to explore experimental design in presence of lineage information, and to connect models with analysis and control of single-cell behavior, also from an experimental viewpoint. The methods developed will be accompanied as much as possible by practical software tools to the profit of the community. Moreover, we intend to pursue an in-depth mathematical analysis of the modelling and estimation approaches developed in [72] as a general advance of the theory of statistics and possible applications to branching process models of different sorts [123].

In the same context, I will also be investigating stochastic gene regulatory network reconstruction from time profiles of gene expression statistics (typically, second-order moments, or population-snapshot histograms). Relying on gene expression statistics instead of single-cell profiles allows one to pool together results from different experiments in a natural fashion, since network statistics are conserved across experiments differing only in the variables observed, and carries potential to outperform deterministic, population-average approaches. My recent contributions [22, 21] constitute the premises of this research line. While focused on reconstruction of biological networks, the work is expected to stimulate more general research on the reconstruction of stochastic dynamical network models from distributed statistical observations, with potential applications *e.g.* to power networks or multi-agent systems [17]. All of these objectives are not only at the core of project MEMIP but also constitute the basis of project follow-ups and novel collaborations with methodological and experimental groups.

Finally, the case studies of Chapter 3 only provide a small set of examples for a variety of relevant reconstruction problems of time-varying cellular activity. The methods developed in the referred papers are powerful tools of much broader applicability. In the context of project MAXIMIC that I participate in ([74], see also below), together with Ph.D. student Antrea Pavlou and her supervisors Hidde de Jong and Johannes Geiselmann, the problem of deconvolving the effects of protein maturation from experimental population and single-cell fluorescent reporter time profiles will be addressed. In parallel, the problem of optimal reporter design for single-cell gene expression measurements will be investigated based on the spectral approach outlined in the chapter, thus marrying the efforts of the MEMIP and MAXIMIC projects. An experimental validation of the spectral design approach is indeed foreseen, within IBIS or perhaps in collaboration with InBio and MSC. Moreover, further applications of the rate estimation method of Section 3.2 and the energetic functioning of the cell will be explored in the context of project RIBECO [101], in particular, with Delphine Ropers (IBIS) and Muriel Cocaign-Bousquet (LISPB, [65]). The project is just starting and includes my participation. As a technological side of the project, development of a user-friendly version of the software implementing the algorithms of [24] is foreseen.

Optimal control of biological systems. We have illustrated in Chapter 4 the relevance of controlling microbial systems for both biological discovery and potential applications of great societal interest. This recent line of my research is currently developed by the support of two projects.

A first project that I participe in, MAXIMIC [74], is devoted to the study of cellular resource allocation strategies and their synthetic re-engineering toward optimal bioproduction of target molecules. Headed by our project-team IBIS [51], it focuses on homogeneous single-strain cellular populations. In this project, besides the estimation questions already mentioned above, nontrivial questions reside not only in the experimental aspects of genetic re-engineering, monitoring and control. Methodological challenges accompany the experimental endeavor, such as the mathematical characterization of the engineered strain, the definition and analysis of relevant optmal control problems, and their solution toward model-based optimal controller design. I will participate to these methodological challenges, notably in collaboration with BIOCORE (Jean-Luc Gouzé, [12]) and the other members of IBIS.

A much related but broader project that I coordinate is the Inria project COSY [29]. Motivated by both the occurrence in natural environments and the possibility for exploitation (see discussions of Chapter 4), we are currently assisting at a real rush toward analysis, synthesis and control of microbial communities [59, 46, 104, 7]. At the same time, given the current technological capabilities, biology is transitioning with impetus toward automation of biological experiments, including the development of on-demand robotized experimental services [106, 113]. In this international context, regrouping five Inria project-teams plus two external partners in a highly interdisciplinary consortium, project COSY addresses automated control of synthetic microbial communities. Optimal production of target biomolecules is one of the application objectives of the project, however, a broader set of optimal control problems constitute the scope of the project. Conceived around the experimental facilities of IBIS (with the experimental team BIOP at LIPHY [13]) and InBio, these also include control of coexistence and size of different subpopulations (see Section 4.3), as well as stochastic control of differentiation and metabolic load at a single-cell or population-snapshot level (see for instance the preliminary results in [121]). This is a fundamental research project meant to be developed at a lab-level, yet, it has potential for real-world applications in the biotech industry. In addition, with the INRA team MaIAGE (Béatrice Laroche) [70], the project aims at an eventual transfer of knowledge from the laboratory study of synthetic microbial communities to the investigation of natural microbial communities [94].

In addition to general coordination, my first scientific contribution to the project is the modelling, analysis and optimal control of a synthetic consortium of *E.coli* bacteria. The details and motivation for the specific consortium we are building across IBIS and BIOP were explained in Section 4.3, where the first modelling and analysis results were anticipated. On these bases, analysis and design of real-time observers and optimal control strategies are being pursued with Inria groups BIOCORE and VALSE (Denis Efimov, [115]). In a new Ph.D. project that will start in the fall 2019 under my co-supervision (with Johannes Geiselmann), Ph.D. student Maaike Sangster will work on

the biological construction of the strains, the experimental analysis and calibration of the models developed in [73], and the deployment of the feedback control strategies developed by the other project-teams of the consortium. Offspring projects from COSY with the same as well as other national and international partners are expected.

A crucial aspect of the COSY project, with impact on all experimental activities including *e.g.* the aforementioned project MAXIMIC, is automation. Time-course monitoring and, more so, real-time control experiments require unceasing laboratory operations on the cultured microbial population. However, the largest part of these operations are routinary. In the interest of repeatability, tracking, parallelization, and (where needed) real-time computational performance, they are ideally best performed by a computer, thus freeing invaluable human capital from tedious, uninteresting activities. In IBIS, we are developing an automated experimental platform consisting of several independently-controlled mini-bioreactors sharing common (costly) measurement devices such as spectrometers and a cytometer. The system is based on a fluidic setup entirely operated by a computer for the transfer of biological samples from bioreactors to the measurement devices and for the actuation of bioreactor control. Currently allowing for injection of liquid solutions carrying biochemical actuators, bioreactor control is being endowed with a LED system for use with optogenetics (gene expression control via light stimuli), the current state of the art in cell control at a lab-level [102, 18]. Started before project COSY, the physical development of this already functional platform is continuing. A first working version of the software architecture for the computer control of the physical platform has been developed in the technological development project COSOFT [28] under my direction. Developed in connection with ODIN, a related software developed at BIOCORE under the direction of Olivier Bernard [82], the software provides a first abstraction layer for the execution of low-level operations and implements scheduling for the utilization of shared resources. In the follow-up technological development project OPTICO that is about to start [83], leveraging the control-theoretic results from COSY, I will direct the development of an interface for the deployment of microbial community control strategies. Both COSOFT and OPTICO profited from the contribution of the development engineer Tamas Muszbek. Developments will constitute the object of technological transfer, to other members of the COSY consortium (a similar platform is being developed at InBio), and beyond.

In summary, automated control of biological experiments in general, and of microbial community dynamics in particular, is a prominent activity at the forefront of current research. As apparent from the description above, the activity entails challenges at all levels (methodological, experimental, and technological) and may thus contribute to several research fields. It will shape the activity of IBIS and of a relevant portion of my research in the future years.

Acknowledgements

The work presented in the manuscript is the result of collaboration and interaction with many excellent researchers. I wish to warmly thank them all. Special thanks go to Giorgio Picci, John Lygeros, Giancarlo Ferrari-Trecate and Hidde de Jong, for providing me with invaluable guidance and inspiration through the different stages of my career, and to all students and young researchers, in particular, Sara Berthomieux, Diana Stefan, Andres Gonzalez-Vargas, Alfonso Carta, Stefano Casagranda, Tamas Muszbek, Aline Marguet, Marco Mauri, Marianna Rapsomaniki, that I had the honor and pleasure to supervise or collaborate with. Thanks to Delphine Ropers, Hans Geiselmann, Thibault Etienne, Antrea Pavlou and all other colleagues at Inria for enriching every day of my research adventure. Thank you Lucia for having supported me at all times of the development of my career. So many more people I wish to thank who entered my working life way beyond the professional level, you have no idea how grateful I am.

Napoli, June 28, 2019

Bibliography

[1] D.K. Agrawal, E. Franco, and R. Schulman. A self-regulating biomolecular comparator for processing oscillatory signals. *Journal of The Royal Society Interface*, 12(111):20150586, 2015.

[2] B.D.O. Anderson and J.B. Moore. *Optimal Control: Linear Quadratic Methods*. Prentice Hall, Upper Saddle River, NJ, 1990.

[3] F. Annunziata, A. Matyjaszkiewicz, G. Fiore, C.S. Grierson, L. Marucci, M. di Bernardo, and N.J. Savery. An orthogonal multi-input integration system to control gene expression in *Escherichia coli*. *ACS Synthetic Biology*, 6(10):1816–1824, 2017.

[4] M. Ashyraliyev, Y. Fomekong Nanfack, J.A. Kaandorp, and J.G. Blom. Systems biology: Parameter estimation for biochemical models. *FEBS J.*, 276(4):886–902, 2009.

[5] K.J. Astrom. *Introduction to Stochastic Control Theory*. Academic Press, New York, 1970.

[6] S..R. Bailey and M.V. Maus. Gene editing for immune cell therapies. *Nature Biotechnology*, 2019.

[7] H.C. Bernstein, S.D. Paulson, and R.P. Carlson. Synthetic *Escherichia coli* consortia engineered for syntrophy demonstrate enhanced biomass productivity. *Journal of Biotechnology*, 157:159–166, 2012.

[8] M. Bertero. Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120, 1989.

[9] S. Berthoumieux, M. Brilli, H. de Jong, D. Kahn, and E. Cinquemani. Identification of linlog models of metabolic networks from incomplete high-throughput datasets. *Bioinformatics (Proceedings of the ISMB conference 2011)*, 27(13):i186–i195, 2011.

[10] S. Berthoumieux, M. Brilli, D. Kahn, H. de Jong, and E. Cinquemani. On the identifiability of metabolic network models. *Journal of Mathematical Biology*, 67(6-7):1795–832, 2013.

[11] S. Berthoumieux, H. de Jong, G. Baptist, C. Pinel, C. Ranquet, D. Ropers, and J. Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular Systems Biology*, 9:634, 2013.

[12] *BIOCORE*. Project-Team, Inria Sophia-Antipolis – Méditerranée. url: `team.inria.fr/biocore`.

[13] *BIOP*. Research team, Laboratoire Interdisciplinaire de Physique (LiPhy), Université Grenoble – Alpes. url: `www-liphy.ujf-grenoble.fr/Equipe-BIOP-presentation`.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, USA, 2004.

[15] C. Briat, A. Gupta, and M.H. Khammash. Antithetic proportional-integral feedback for reduced variance and improved control performance of stochastic reaction networks. *Journal of the Royal Society Interface*, 15(143):20180079, 2018.

[16] M. Campi and S. Garatti. *Introduction to the Scenario Approach*. SIAM, U.S., 2019.

[17] Y. Cao, W. Yu, W. Ren, and G. Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1):427–438, 2013.

[18] R. Chait, J. Ruess, T. Bergmiller, G. Tkacik, and C.C. Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature Communications*, 8(1):1535, 2017.

[19] O.T. Chis, J.R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: a critical comparison of methods. *PLoS One*, 6(11):e27755, 2011.

[20] K.H. Cho, S.M. Choo, S.H. Jung, J.R. Kim, H.S. Choi, and J. Kim. Reverse engineering of gene regulatory networks. *IET Systems Biology*, 1(3):149–63, 2007.

[21] E. Cinquemani. Identifiability and reconstruction of biochemical reaction networks from population snapshot data. *Processes (Special Issue on Computational Synthetic Biology)*, 6(9):136, 2018.

[22] E. Cinquemani. Stochastic reaction networks with input processes: Analysis and application to gene expression inference. *Automatica*, 101:150–156, 2019.

[23] E. Cinquemani, M. Agarwal, D. Chatterjee, and J. Lygeros. Convexity and convex approximations of discrete-time stochastic control problems with constraints. *Automatica*, 47:2082–87, 2011.

[24] E. Cinquemani, V. Laroute, M. Cocaign-Bousquet, H. de Jong, and D. Ropers. Estimation of time-varying growth, uptake and excretion rates from dynamic metabolomics data. *Bioinformatics (Proceedings of ISMB/ECCB conference 2017)*, 33:i301–i310, 2017.

[25] E. Cinquemani, F. Mairet, I. Yegorov, H. de Jong, and J.-L. Gouzé. Optimal control of bacterial growth for metabolite production: The role of timing and costs of control. In *Proceedings of the 17th European Control Conference (ECC 2019)*, 2019.

[26] A. Colman-Lerner, A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C.G. Pesce, and R. Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437:699–706, 2005.

[27] *COMMANDS*. Project-Team, Inria Saclay – Île-de-France. url: `team.inria.fr/commands`.

[28] COSOFT – Control software for a system of mini-bioreactors. Inria Action de Développement Technologique (ADT), 2017.

[29] COSY – Real-time control of synthetic microbial communities. Inria Project-Lab (IPL), 2017–2021. url: `project.inria.fr/iplcosy`.

[30] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[31] H. de Jong, S. Casagranda, N. Giordano, E. Cinquemani, D. Ropers, J. Geiselmann, and J.-L. Gouzé. Mathematical modeling of microbes: Metabolism, gene expression, and growth. *Journal of the Royal Society Interface*, 14:20170502, 2017.

[32] H. de Jong, C. Ranquet, D. Ropers, C. Pinel, and J. Geiselmann. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Systems Biology*, 4(1):55, 2010.

[33] G. De Nicolao, G. Sparacino, and C. Cobelli. Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica*, 33(5):851–70, 1997.

[34] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

[35] A. Doucet, A. Smith, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[36] DREAM – Dialogue for Reverse Engineering Assessments and Methods. Initiative. url: `http://dreamchallenges.org`.

[37] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[38] S. Estrela, C.H. Trisos, and S.P. Brown. From metabolism to ecology: Cross-feeding interactions shape the balance between polymicrobial conflict and mutualism. *The American Naturalist*, 180(5):566–76, 2012.

[39] K. Faust, F. Bauchinger, B. Laroche, S. de Buyl, L. Lahti, A. D. Washburne, D. Gonze, and S. Widder. Signatures of ecological processes in microbial community time series. *Microbiome*, 6:120, 2018.

[40] B. Finkenstädt, E.A. Heron, M.l Komorowski, K. Edwards, S. Tang, C.V. Harper, J.R.E. Davis, M.R.H. White, A.J. Millar, and D.A. Rand. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901–2907, 2008.

[41] F. Fröhlich, A. Reiser, L. Fink, D. Woschée, T. Ligon, F.J. Theis, J.O. Rädler, and J. Hasenauer. Multi-experiment nonlinear mixed effect modeling of single-cell translation kinetics after transfection. *NPJ Systems Biology and Applications*, 5(1), 2018.

[42] D. T. Gillespie. The chemical langevin equation. *Journal of Chemical Physics*, 113:297–306, 2000.

[43] D.T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425, 1992.

[44] N. Giordano, F. Mairet, J.L. Gouzé, J. Geiselmann, and H. de Jong. Dynamical allocation of cellular resources as an optimal control problem: Novel insights into microbial growth strategies. *PLoS Computational Biology*, 12(3):e1004802, 2016.

[45] J. Grefenstette, S. Kim, and S. Kauffman. An analysis of the class of gene regulatory functions implied by a biochemical model. *Biosystems*, 84(2):81–90, 2006.

[46] E. Harvey, J. Heys, and T. Gedeon. Quantifying the effects of the division of labor in metabolic pathways. *Journal of Theoretical Biology*, 360:222–242, 2014.

[47] J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgower. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(1):125, 2011.

[48] J.P. Hespanha. Modelling and analysis of stochastic hybrid systems. *IEE Proceedings – Control Theory and Applications*, 153(5):520–535, Sept 2006.

[49] J. Hesseler, J.K. Schmidt, U. Reichl, and D. Flockerzi. Coexistence in the chemostat as a result of metabolic by-products. *Journal of Mathematical Biology*, 53:556–584, 2006.

[50] A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *PNAS*, 108(29):12167–12172, 2011.

[51] *IBIS*. Project-Team, Inria Grenoble – Rhône-Alpes. url: `team.inria.fr/ibis`.

[52] *INBIO*. Exploratory action, Inria Saclay – Île-de-France / Institut Pasteur. url: `research.pasteur.fr/en/team/experimental-and-computational-methods-for-modeling-cellular-processes`.

[53] N. Ishii *et al*. Multiple high-throughput analyses monitor the response of *E.coli* to perturbations. *Science*, 316(5824):593–597, 2007.

[54] J. Izard, C. Gomez Balderas, D. Ropers, S. Lacour, X. Song, Y. Yang, A.B. Lindner, J. Geiselmann, and de Jong. H. A synthetic growth switch based on controlled expression of rna polymerase. *Molecular Systems Biology*, 11(11):840, 2015.

[55] S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–22, Mar 2004.

[56] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[57] M. Komorowski, B. Finkenstädt, C. Harper, and D. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1):343, 2009.

[58] O. Kotte, B. Volkmer, J.L. Radzikowski, and M. Heinemann. Phenotypic bistability in *Escherichia coli*'s central carbon metabolism. *Molecular Systems Biology*, 10:736–736, 2014.

[59] J.U. Kreft, C.M. Plugge, C. Prats, J.H.J. Leveau, W. Zhang, and F.L. Hellweger. From genes to ecosystems in microbiology: Modeling approaches and the importance of individuality. *Frontiers in Microbiology*, 8(2299), 2017.

[60] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling. Ensemble modeling for analysis of cell signaling dynamics. *Nature Biotechnology*, 25(9):1001–6, 2007.

[61] M. Lavielle. *Mixed effects models for the population approach. Models, Tasks, Methods & Tools.* Chapman & Hall/CRC Biostatistics Series. CRC press, 2015.

[62] I. Lestas, J. Paulsson, N. E. Ross, and G. Vinnicombe. Noise in gene regulatory networks. *IEEE Transactions on Automatic Control*, 53(Special Issue):189–200, 2008.

[63] G. Lillacci, Y. Benenson, and M.H. Khammash. Synthetic control systems for high performance gene expression in mammalian cells. *Nucleic Acids Research*, 46(18):9855–9863, 2018.

[64] G. Lillacci and M. Khammash. Parameter estimation and model selection in computational biology. *PLoS Computational Biology*, 6(3):e1000696, 2010.

[65] *LISPB*. INSA laboratory, Toulouse. url: `www.lisbp.fr`.

[66] L. Ljung. *System Identification – Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 1999.

[67] A. Llamosi, A.M. Gonzalez-Vargas, C. Versari, E. Cinquemani, G. Ferrari-Trecate, P. Hersen, and G. Batt. What population reveals about individual cell identity: Single-cell parameter estimation of models of gene expression in yeast. *PLOS Computational Biology*, 12(2), 2016. Available online at https://doi.org/10.1371/journal.pcbi.1004706.

[68] J. Lygeros and M. Prandini. Stochastic hybrid systems: A powerful framework for complex, large scale applications. *European Journal of Control*, 16:583–594, 2010.

[69] M.L. Maeder and C.A. Gersbach. Genome-editing technologies for gene and cell therapy. *Molecular Therapy*, 24(3):430–46, 2016.

[70] *MaiAGE*. Research unit, INRA Jouy-en-Josas. url: `http://maiage.jouy.inra.fr/`.

[71] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, The DREAM5 Consortium, M. Kellis, J.J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796–804, 2012.

[72] A. Marguet, M. Lavielle, and E. Cinquemani. Inheritance and variability of kinetic gene expression parameters in microbial cells: Modelling and inference from lineage tree data. *Bioinformatics (Proceedings of the ISMB/ECCB conference 2019)*, To appear., 2019.

[73] M. Mauri, J.L. Gouzé, H. de Jong, and E. Cinquemani. Modelling and design of a synthetic microbial community for enhanced productivity in bioreactor. *In preparation*, 2019.

[74] MAXIMIC – Optimal control of microbial cells by natural and synthetic strategies. ANR project (ANR-17-CE40-0024), 2017–2021. url: `anr.fr/Project-ANR-17-CE40-0024`.

[75] MEMIP – Mixed-Effects models of intracellular processes: Methods, tools and applications. ANR project (ANR-16-CE33-0018), 2016–2020. url: `anr.fr/Projet-ANR-16-CE33-0018`.

[76] A. Milias-Argeitis, M. Rullan, S.K. Aoki, P. Buchmann, and M.H. Khammash. Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth. *Nature Communications*, 7:12546, 2016.

[77] A. Milias-Argeitis, S. Summers, J. Stewart-Ornstein, I. Zuleta, D. Pincus, H. El-Samad, M.H. Khammash, and J. Lygeros. *In silico* feedback for *in vivo* regulation of a gene expression circuit. *Nature Biotechnology*, 29:1114–1116, 2011.

[78] *MSC* (laboratoire matière et systèmes complexes). CNRS / Université Paris-Diderot. url: `www.msc.univ-paris-diderot.fr`.

[79] B. Munsky, B. Trinh, and M. Khammash. Listening to the noise: Random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(318), 2009.

[80] N. Nakamura, H.C. Lin, C.S. McSweeney, R.I. Mackie, and H.R. Gaskins. Mechanisms of microbial hydrogen disposal in the human colon and implications for health and disease. *Annual Review of Food and Science Technology*, 1:363–395, 2010.

[81] G. Neuert, B. Munsky, R.Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, 2013.

[82] ODIN. Inria Action de Développement Technologique (ADT), 2012. url: `team.inria.fr/biocore/software/odin`.

[83] OPTICO – Optimal control software for microbial communities in a system of mini-bioreactors. Inria Action de Développement Technologique (ADT), 2019.

[84] J.D. Orth, I.Thiele, and B.O. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

[85] D.A. Oyarzún and M. Chaves. Design of a bistable switch to control cellular uptake. *Journal of the Royal Society Interface*, 12(113):20150618, 2015.

[86] G.J. Patti, O. Yanes, and G. Siuzdak. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–9, 2012.

[87] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157 – 175, 2005.

[88] J.M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307:1965–1969, 2005.

[89] R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate. Identification of genetic network dynamics with unate structure. *Bioinformatics*, 26(9):1239–45, 2010.

[90] R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate. Invalidation of the structure of genetic network dynamics: A geometric approach. *International Journal of Robust and Nonlinear Control (Special Issue on System Identification for Biological Systems)*, 22(10):1140–56, 2012.

[91] L. Postiglione, S. Napolitano, E. Pedone, D.L. Rocca, F. Aulicino, M. Santorelli, B. Tumaini, L. Marucci, and D. di Bernardo. Regulation of gene expression and signaling pathway activity in mammalian cells by automated microfluidics feedback control. *ACS Synthetic Biology*, 7(11):2558–2565, 2018.

[92] A. Prékopa. *Stochastic Programming*, volume 324 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995.

[93] J. Qin *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65, 2010.

[94] S. Raguideau, S. Plancade, N. Pons, M. Leclerc, and B. Laroche. Inferring aggregated functional traits from metagenomic data using constrained non-negative matrix factorization: Application to fiber degradation in the human gut microbiota. *Plos computational Biology*, 12(12):1–29, 2016.

[95] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135:216–226, 2008.

[96] M.A. Rapsomaniki, E. Cinquemani, N.N. Giakoumakis, P. Kotsantis, J. Lygeros, and Z. Lygerou. Inference of protein kinetics by stochastic modeling and simulation of fluorescence recovery after photobleaching experiments. *Bioinformatics*, 31(3):355–362, 2014.

[97] J.M. Raser and E.K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811–1814, 2004.

[98] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006.

[99] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–29, 2009.

[100] J. B. Rawlings and D. Q. Mayne. *Model Predictive Control: Theory and Design*. Nob Hill Publishing, LLC, 2009.

[101] RIB-ECO – Engineering RNA life cycle to optimize economy of microbial energy: Application to the bioconversion of biomass-derived carbon sources. ANR project (ANR-18-CE43-0010), 2018–2022. url: `project.inria.fr/ribeco`.

[102] M. Rullan, D. Benzinger, G. Schmidt, A. Milias-Argeitis, and M.H. Khammash. An optogenetic platform for real-time, single-cell interrogation of stochastic transcriptional regulation. *Molecular Cell*, 70(4):745–756.e6, 2018.

[103] A. Samson, M. Lavielle, and F. Mentré. Extension of the SAEM algorithm to left-censored data in non-linear mixed-effects model: Application to hiv dynamics model. *Computational Statistics and Data Analysis*, 51(3):1562–74, 2006.

[104] S. Santala, M. Karp, and V. Santala. Rationally engineered synthetic coculture for improved biomass and product formation. *PLoS One*, 9(12):e113786, 2014.

[105] M. Schelker, A. Raue, J. Timmer, and C. Kreutz. Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinformatics*, 28(18):i529–34, 2012.

[106] M. Scudellari. Software startups aim to automate bio labs. *IEEE spectrum*, 2017. url: `spectrum.ieee.org/the-human-os/biomedical/devices/software-startups-aim-to-automate-bio-labs`.

[107] D. Stefan, C. Pinel, S. Pinhal, E. Cinquemani, J. Geiselmann, and H. de Jong. Inference of quantitative models of bacterial promoters from time-series reporter gene data. *PLoS Comput. Biol.*, 11(1):e1004028, 01 2015.

[108] A. Succurro, D. Segrè, and O. Ebenhöh. Emergent subpopulation behavior uncovered with a community dynamic metabolic model of *Escherichia coli* diauxic growth. *mSystems*, 4(1), 2019.

[109] P.S. Swain, M.B. Elowitz, and E.D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20):12795–12800, 2002.

[110] Y. Taniguchi, P.J. Choi, G.W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X.S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, 2010.

[111] O. Tenaillon, D. Skurnik, Picard B., and E. Denamur. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology*, 8(3):207–17, 2010.

[112] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *PNAS*, 98(15):8614–8619, 2001.

[113] *Transcriptic Inc.* Menlo Park, CA. url: `www.transcriptic.com`.

[114] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. *PNAS*, 109(35):14271–14276, 2012.

[115] *VALSE*. Project-Team, Inria Lille – Nord Europe. url: `team.inria.fr/valse`.

[116] D. Del Vecchio, A.J. Dy, and Y. Qian. Control theory meets synthetic biology. *Journal of The Royal Society Interface*, 13(120):20160380, 2016.

[117] A. Villaverde and J. Banga. Reverse engineering and identification in systems biology: Strategies, perspectives and challenges. *Journal of the Royal Society Interface*, 11:20130505, 2014.

[118] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

[119] S. Waldherr. Estimation methods for heterogeneous cell population models in systems biology. *Journal of The Royal Society Interface*, 15(147):20180530, 2018.

[120] L. Weber, W. Raymond, and B. Munsky. Identification of gene regulation models from single-cell data. *Physical Biology*, 15(5):055001, 2018.

[121] E. Weill, V. Andreani, C. Aditya, P. Martinon, G. Batt, and F. Bonnans. Optimal control of an artificial microbial differentiation system for protein bioproduction. In *Proceedings of the 17th European Control Conference (ECC 2019)*, 2019.

[122] S. Widder *et al*. Challenges in microbial ecology: Building predictive understanding of community function and dynamics. *The ISME Journal*, 10:2557–2568, 2016.

[123] A. S. Willsky. Multiresolution M]arkov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002.

[124] *XPOP*. Project-Team Inria Saclay – Île-de-France / CMAP École Polytechnique. url: `www.inria.fr/en/teams/xpop`.

[125] I. Yegorov, F. Mairet, H. de Jong, and J.-L. Gouzé. Optimal control of bacterial growth for the maximization of metabolite production. *Journal of Mathematical Biology*, 2018.

[126] E. Yeung, A.J. Dy, K.B. Martin, A.H. Ng, D. Del Vecchio, J.L. Beck, J.J. Collins, and R.M. Murray. Biophysical constraints arising from compositional context in synthetic gene networks. *Cell Systems*, 5(1):11–24, 2017.

[127] H. Youk and W.A. Lim. Secreting and sensing the same molecule allows cells to achieve versatile social behaviors. *Science*, 343:l1242782, 2014.

[128] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *PNAS*, 109(21):8340–8345, 2012.

[129] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11:197–202, 2014.

[130] V. Zulkower, M. Page, D. Ropers, J. Geiselmann, and H. de Jong. Robust reconstruction of gene expression profiles from reporter gene data using linear inversion. *Bioinformatics*, 31(12):i71–9, 2015.